Flinders University

Machine Learning and Data Mining for Sports Performance Analytics

Frameworks & Applications for Match Outcome Prediction and Pattern Identification

By

Rory Paul Bunker
Bachelor of Commerce (Honours), Master of Analytics

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy (PhD) by Prior Published Work

College of Science and Engineering
7 April 2025

Principal Supervisor: Dr. Matthew Stephenson, College of Science and Engineering, Flinders University

Associate Supervisor: Dr. Kym Williams, Sport and Active Recreation College of Education, Psychology and Social Work, Flinders University

TABLE OF CONTENTS

ABSTRACT

Identifying the factors contributing to successful outcomes at different levels is vital for decision-makers in professional sports, which historically have relied on subjective opinion. Technological developments have resulted in rapid growth in data generated in sports, which, with appropriate techniques, ostensibly allows for the objective identification of factors contributing to success.

This thesis presents original contributions to the emerging field of sports performance analytics through six interconnected publications addressing three primary research questions. The first question investigates how conceptual frameworks can be developed for machine learning for match outcome prediction and sports performance analytics as a whole. The second question explores how data mining methods from other disciplines can be effectively leveraged for key pattern identification in sports. The third question focuses on applying machine learning methods from various domains to provide interpretable match outcome predictions in sports.

The thesis highlights the benefits of contemporary machine learning and data mining approaches used successfully in other domains and contexts for match outcome prediction and key pattern identification. In this thesis, relevant literature in machine learning for sports match outcome prediction is critically analysed and synthesised, and a conceptual framework for applying machine learning in sports match result prediction is proposed. This framework is then demonstrated in practice in a subsequent publication in the context of tennis match result prediction. More specific explorative surveys into machine learning for match outcome prediction were carried out in the context of team sports in general and then soccer specifically.

Sequential pattern mining-based classifiers are used to identify interpretable key event patterns from passages of play that discriminate between scoring and non-scoring outcomes, and interpretable rules-based machine learning is used to identify key patterns composed of performance indicator combinations and values that discriminate between successful and unsuccessful tournament stage outcomes. This thesis takes a multi-level approach, investigating performance at various levels of analysis, including the match outcome, tournament stage, and passages of play levels across various sports.

DECLARATIONS

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university
2. and the research within will not be submitted for any other future degree or diploma without the permission of Flinders University
3. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text, and
4. if generative artificial intelligence has been used in my thesis it has been duly acknowledged with details to identify the extent to which generative artificial intelligence formed the final thesis

Rory P. Bunker

Date.....4 November, 2024.....

Peer review

All of the publications included in this PhD by prior publication thesis have been peer-reviewed and approved by the editors of the journals according to their publication standards and the standards of the discipline wherein they are published.

Declaration of Ethics

Ethics approval was not required for any of the publications included in this thesis.

Contribution Declaration

I declare that for all publications included in this thesis:

- I was the first and primary author to contribute to the conception, investigation, and manuscript draft writing. I was also the main person who edited all manuscripts before submission and handled the re-submission and publication process as the corresponding author.
- I singly conceived the studies and sought support from co-authors, whom I had become acquainted with or worked with during my employment.
- I conducted the data analysis and modelling in all applicable publications.

Note: *The services of a professional editor were not used in preparing this contextual statement nor in any of the publications in this thesis.*

ACKNOWLEDGEMENTS

Funding: *I wish to acknowledge the contribution provided by the Australian Government Research Training Program Scholarship (fees offset).*

A PhD is a significant undertaking but an important step in terms of the training it provides to think like a scientist and researcher and to open the pathway to pursuing an academic or research career. I took an unusual — and perhaps long-winded — route to writing a PhD by prior publication, and I was grateful for the opportunity to pursue this qualification through Flinders University.

I'd like to acknowledge several people for their support over the past several years. First, I would like to acknowledge my prior supervisors, Keisuke Fujii and Ichiro Takeuchi, whom I reported to at Nagoya University and Nagoya Institute of Technology. I developed my research skills greatly under their tutelage. Your support and knowledge are greatly appreciated.

I would also like to thank my principal PhD supervisor at Flinders University, Matthew Stephenson, who took on this supervisory role when my original principal supervisor, Kym Williams, went on parental leave. I am also grateful to Kym, who was very supportive in our initial meetings, assisted in commencing the enrolment period, and remained an associate supervisor. Your support and guidance are very much appreciated.

Also, I would like to thank all of the co-authors who have contributed to this thesis's publications. In particular, I would like to thank Kirsten Spencer from the Sports Performance Research Institute New Zealand at Auckland University of Technology, Teo Susnjak from the School of Mathematical and Computational Sciences at Massey University, Fadi Thabtah from Digital Technologies at Manukau Institute of Technology, and Calvin Yeung, Ichiro Takeuchi, and Keisuke Fujii from Nagoya University in Japan.

Finally, I am immensely grateful for the support of my family, including that of my wife, my two sons, and my wider family.

PERSONAL INTRODUCTION

This section briefly introduces me and discusses some of my background. In particular, it outlines the events that led to the conceptualisation and eventual acceptance/publication of the studies that make up this thesis and describes my personal research employment history. In doing so, this section provides background in terms of how the research originated and evolved throughout my employment/research experience in recent years.

Research Origins: Master of Analytics Research Methods Course & Performance Analyst Work Experience (start of 2015 – mid-2016)

While studying for a Master of Analytics at Auckland University of Technology (AUT) in 2015, I worked part-time as a performance/video analyst at Counties Manukau Rugby Union, a professional rugby team located in the Franklin and South Auckland regions of New Zealand. This work involved sourcing data from centralised Opta/New Zealand Rugby databases, which are available to all teams in the National Provincial Championship (NPC) rugby competition, developing reports to monitor own and opposition player- and team-level performance for coaches, filming of matches, and importing match video into SportsCode for event tagging.

A research methods course was one of the first-semester core course requirements of the Master of Analytics. It involved working on a small research project that utilised some of the research methods learnt during the course. I undertook a project that used the type of data that had been captured and utilised in the course of the work with Counties Rugby, and I submitted a research report for the course on the topic of machine learning for rugby union match result prediction.

Continuation of Research: A chance encounter & continuation post-graduation out of personal interest (mid 2016 – 2018)

After completing my Masters degree in July 2016, I took a trip to Japan (and Beijing, China for a few days) with my girlfriend (now wife). While applying for a tourist visa at the Chinese Embassy in Auckland, my car was unfortunately clamped. As I was discussing the firm's questionable clamping practices with the security guard, a bystander, Fadi Thabtah, sympathised with my predicament. Fadi and I began talking about work matters. It turned out that he was an academic conducting research in machine learning at Nelson-Marlborough Institute of Technology in Auckland.

A few weeks later, I emailed Fadi about the machine learning for sports result prediction project I had worked on at AUT, which I had recently started to think about again. I visited him at his office, and we discussed the project further. Later, we communicated about the research via email. The research shifted to summarising the literature on sports match result prediction using artificial neural networks (ANNs) and establishing a conceptual framework for sports match result prediction using machine learning (Bunker & Thabtah, 2019).

Near the end of 2016, I contemplated beginning a PhD in Nagoya, Japan, my girlfriend's (now wife's) home city. I contacted Professor Ichiro Takeuchi from Nagoya Institute of Technology, who was attending the 8th Asian Conference on Machine Learning at the University of Waikato in November 2016. I arranged to meet with him at Auckland Airport before his flight home, and we discussed common research interests, his laboratory's specialities, and a potential PhD topic. Since I took on a role in data analytics in industry, I did not enrol in a PhD at this time. However, I would contact Professor Takeuchi again in the not-too-distant future.

Research employment at universities in Japan (2019-2024)

My wife and I relocated to Nagoya in February 2019. Before the move, I emailed Professor Takeuchi to see if he had any work available in his laboratory at Nagoya Institute of Technology. Professor Takeuchi's laboratory, the Takeuchi-Karasuyama Laboratory (now the Karasuyama-Inatsu Laboratory) at Nagoya Institute of Technology, focused on machine learning research and data science practice, with applications in domains such as bioinformatics, medical informatics, and materials informatics. In October 2019, I attended the Asia-Pacific Conference on Performance Analysis of Sports at Nagoya University. I obtained accreditation from the International Society of Performance Analysis of Sport (ISPAS) Level 1 performance analyst.

In Professor Takeuchi's lab, I initially worked on identifying allergen subsequences from labelled protein sequence data using a discriminative sequential pattern mining technique that lab members had developed. Given my continuing interest in sports, we contemplated whether this technique could be applied to sports event sequences (passages of play). For this, I sourced a study dataset from a professional rugby team in Japan. This led to a publication (Bunker, Fujii, Hanada, & Takeuchi, 2021), which appears in this thesis, and a conference proceedings paper and presentation (Bunker, 2022).

I moved from the Takeuchi-Karasuyama Laboratory to the Sports Behavior Group within the Takeda Behavioral Signal Processing Laboratory at Nagoya University at the start of April 2021 as it closely aligned with my research interests. I worked under the supervision of Dr. Keisuke Fujii, one of the publication's co-authors (Bunker et al., 2021). During this time, I attained the ISPAS Level 2 Performance Analyst accreditation and ISPAS Level 3 Performance Analyst accreditation (Scientific route), the latter based on the papers I had published in the domain. I still maintain the ISPAS Level 3 Performance Analyst accreditation under the scientific route when writing this contextual statement.

Initial contact with Flinders University about the PhD by prior publication

As I produced more publications, I investigated the possibility of gaining recognition for this existing work I had already completed rather than starting a PhD from scratch. The Flinders University PhD by prior published work program provided the opportunity to do this. I contacted Dr. Kym Williams, who had research interests in a different area of sports science but shared a common interest in applying data analytics in sports. After we met virtually, Kym agreed to act as my supervisor for my PhD by prior publication.

Although Kym went on parental leave for a large portion of the enrollment period of my PhD by prior publication in 2024, he was instrumental in the initial steps to enrol in the PhD by prior publication and in finding a new principal supervisor, Dr Matthew Stephenson. Matthew has deep domain knowledge of machine learning, data science, and artificial intelligence in the context of gaming and game analytics. Kym remained an associate supervisor for this PhD by prior publication.

LIST OF ABBREVIATIONS

This list of abbreviations excludes journal names, academic degrees, place names, and where an abbreviation comprises part, or all, of the name of an organisation.

- **ADTree** Alternating Decision Tree
- **AFL** Australian Football League
- **AI** Artificial Intelligence
- **ANOVA** Analysis of Variance
- **ANN** Artificial Neural Network
- **ATP** Association of Tennis Professionals (men's tennis league)
- **AUT** Auckland University of Technology
- **CatBoost** Categorical Boosting (algorithm)
- **CHAID** Chi-squared automatic interaction detection (decision tree)
- **CM-SPADE** Co-occurrence Map Sequential Pattern Discovery using Equivalent Class
- **CM-SPAM** Co-occurrence Map Sequential Pattern Analysis Method
- **CRISP-DM** Cross-Industry Standard Process Data Mining (framework)
- **DOI** Digital Object Identifier
- **GAP** Generalised Attacking Performance (ratings)
- **GSP** Generalized Sequential Pattern (algorithm)
- **ISPAS** International Society of Performance Analysis of Sport
- **KDD** Knowledge Discovery in Databases
- **LSTM** Long short-term memory
- **LSVC** Linear Support Vector Classifier
- **MA-Stat-DSM** Multi-Agent Statistically Discriminative Sub-Trajectory Mining (algorithm)
- **ML** Machine Learning
- **NPC** National Provincial Championship (top-level domestic rugby competition in New Zealand)
- **NRL** National Rugby League
- **PI** Performance Indicator
- **PostgreSQL** Postgres Structured Query Language
- **RIPPER** Repeated Incremental Pruning to Produce Error Reduction (algorithm)

- **RMSE** Root mean squared error
- **RPS** Ranked probability score
- **RQ** Research question
- **RWC** Rugby World Cup
- **SHAP** TreeExplainer SHapley Additive exPlanations (explainable AI method)
- **SPADE** Sequential Pattern Discovery using Equivalent Class (algorithm)
- **SPP** Safe Pattern Pruning (algorithm)
- **SQL** Structured Query Language
- **SRP-CRISP-DM** Sports Result Prediction Cross-Industry Standard Process Data Mining (framework)
- **Stat-DSM** Statistically Discriminative Sub-Trajectory Mining (algorithm)
- **VAEP** Valuing Actions by Estimating Probabilities
- **WElo** Weighted Elo (ratings)
- **WTA** Women's Tennis Association (women’s tennis league)
- **WY** Westfall-Young (method)
- **xAI** Explainable AI
- **xG** Expected goals
- **XGBoost** eXtreme Gradient Boosting (algorithm)
- **XML** Extensible Markup Language

GLOSSARY

A glossary of terms is provided below, some of which are generic descriptions of the term and some of which relate to how the term is used in this contextual statement.

- **Advanced analytics** is a branch of analytics that extends beyond data analysis into computer science- and data science-based analytical techniques that span, for example, machine learning, data mining, and multivariate statistics. Advanced analytics models are sometimes predictive.
- **Boosted tree model:** Tree-based machine learning models that utilise boosting algorithms. This thesis mentions gradient-boosted tree models such as XGBoost and CatBoost, which use the gradient descent algorithm, and a boosted tree model called Alternating Decision Trees that uses another form of boosting, AdaBoost.
- **Coding:** Annotating events in match video using a notational video analysis system.
- **Coding window:** A visual user interface created by a performance analyst within a video analysis system that contains buttons, each used to annotate events of interest from match video.
- **Competitive balance** is a concept from sports economics related to the degree of competitiveness in a particular sport or league.
- **Contextual information/variable:** Information often external to matches, i.e., unrelated to events that occur within matches (e.g., venue, opposition quality, stage of the season), cannot be changed by a team regardless of performance. Sometimes contextual information can also be within a match, for example, the context in terms of whether a player was under pressure by an opposition player while making a pass in soccer/basketball or the position that event occurred on the field. Also known as situational information/variables.
- **Deep learning** is a sub-field of machine learning that considers the use of deep neural networks that possess many layers between the network input and output.
- **Decision rule**: An if-then type statement consisting of a condition and a prediction.
- Discriminative: Discriminative means that the pattern or variable discriminates between data labels.
- **Domain knowledge:** Expert knowledge of a specific field. In this thesis, generally, this means expert knowledge of the sport in question.
- **Dual sport**: A sport that is typically played between two opposing players (e.g., tennis).

- **Event log:** A log containing event data (e.g., for a particular match) that can be exported, for example, in XML format from a notational/video analysis system.
- **Explainable AI:** Methods that allow users to understand the results generated by machine learning (or often deep learning) models.
- **Feature**: A variable in a machine learning model.
- **Feedback loop:** An iterative flow of information between coaches and performance analysts. For instance, the performance analyst relays statistics highlighting team and player performance to coaches; based on this information, coaches adjust what events they want to focus on and have annotated by the analyst, for example, in subsequent post-match reviews.
- **Invasion sports:** Team sports are time-dependent and aim to invade the opposition's territory to score more goals than the opponent within an allocated time frame (e.g., soccer, hockey, rugby).
- **Key pattern:** "Key" in this context means relevant/important/discriminative.
- **Knowledge discovery** is the process of extracting useful knowledge from data. Techniques for knowledge discovery include — but are not limited to — machine learning, data mining, and statistical techniques.
- **Notational system** software is used to annotate events from match videos.
- **Operational event definitions:** Standard definitions of what constitutes a specific event in a particular sport for tagging purposes (e.g., what exactly constitutes a turnover in soccer and how it can be recognised from match film).
- **Optical system:** A setup of arena cameras to track players and the ball and obtain high-frequency spatiotemporal data.
- **PostgreSQL:** An open-source relational database management system.
- **Player group/unit:** A collection of players based on, for example, position or role (e.g., offensive, defensive).
- **Rating system**: Methods for computing ratings representing the relative ability of teams or players.
- **Reliability:** In performance analysis of sports, reliability often refers to inter-operator reliability, the discrepancy between the coded events from two analysts annotating events from the same match (measured using statistical techniques).
- **Situational information/variables:** See contextual information/variable.
- **Sports management analytics** is a sub-field of sports analytics that primarily focuses on the financial and economic aspects of professional sports (as opposed to performance).

- **Sports performance analytics** is a sub-field of sports analytics that primarily focuses on the performance-related aspects of professional sports. This field, this contextual statement proposes, has emerged in recent years due to the increasing use of advanced analytics techniques, which have been traditionally employed in sports analytics and sports performance analysis.
- **Striking/fielding sports:** Team sports that are innings-based and typically involve fielders and striking a ball with some form of bat (e.g., cricket, baseball)
- **Support**: A count representing how often a particular pattern appears in a dataset.
- **Tagging**: See “coding”
- **Target variable:** The variable predicted by a machine learning model (also known as a class variable).
- **(Un)supervised learning:** sub-fields of machine learning involving (un)labelled data.
- **Video analysis system**: See “notational systems”
- **Wearable devices** include Global positioning system (GPS) based devices that players wear to track their movement.

LIST OF FIGURES

LIST OF TABLES

LIST OF PUBLICATIONS INCLUDED IN THIS THESIS & RELATED PUBLICATIONS

Publications included in the main text of the thesis

Six publications are included in the main text of this PhD by prior publication thesis, the details of which are listed below. All of the papers were either already published in academic journals or, in the case of one publication, a book chapter (Bunker, Yeung, & Fujii, 2025), accepted for publication before commencing the enrolment period for the PhD by prior publication.

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied computing and informatics*, 15(1), 27-33. https://doi.org/10.1016/j.aci.2017.09.005

Bunker, R., Fujii, K., Hanada, H., & Takeuchi, I. (2021). Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: an application to rugby union. *PloS one*, 16(9), e0256329. https://doi.org/10.1371/journal.pone.0256329

Bunker, R.P., & Spencer, K. (2022). Performance indicators contributing to success at the group and play-off stages of the 2019 Rugby World Cup. *Journal of Human Sport & Exercise*, 17(3): 683-698. https://doi.org/10.14198/jhse.2022.173.18

Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, 73, 1285-1322. https://doi.org/10.1613/jair.1.13509

Bunker, R., Yeung, C., Susnjak, T., Espie, C., & Fujii, K. (2023). A comparative evaluation of Elo ratings and machine learning-based methods for tennis match result prediction. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, 17543371231212235. https://doi.org/10.1177/17543371231212235

Bunker, R., Yeung, C., Fujii, K. (2025). Machine Learning for Soccer Match Result Prediction. In: Blondin, M.J., Fister Jr., I., Pardalos, P.M. (eds) *Artificial Intelligence,*

Optimization, and Data Sciences in Sports. Springer Optimization and Its Applications, vol 218. Springer, Cham. https://doi.org/10.1007/978-3-031-76047-1_2

Publications included in the appendix to the thesis

The two papers listed below were accepted for publication after commencing the PhD by prior publication enrolment period and, as a result, could not be included in the main text of the thesis. However, these two studies are connected to the six papers in the main text, are mentioned in this contextual statement, and are therefore included in the appendix to this thesis (Appendix A2). Bunker, Yeung, & Fujii (2024) was accepted for publication while enrolled in the PhD by prior publication, and Bunker et al. (2024) was published while enrolled in the PhD by prior publication.

Bunker, R. P., Yeung, C., & Fujii, K. (2024). An expected wins approach using Fisher's Exact Test to identify the bogey effect in sports: An application to tennis. *Journal of Sport & Exercise Science*, 8(1), 43-54. https://doi.org/10.36905/jses.2024.01.06

Bunker, R., Duy, V. N. L., Tabei, Y., Takeuchi, I., & Fujii, K. (2024). Multi-agent statistical discriminative sub-trajectory mining and an application to NBA basketball. *Journal of Quantitative Analysis in Sports*. 2024 Sep 23. https://doi.org/10.1515/jqas-2023-0039

Other connected conference publications/presentations

The following were presented at conferences and are directly related to two of the publications in the thesis's main text and one of the publications in the appendix.

Bunker, R., Fujii, K., Hanada, H., & Takeuchi, I. (2021). Supervised sequential pattern mining for identifying important patterns of play in rugby. In *Proceedings of the 8th MathSport International Conference*, June 2021. Online.

Bunker, R. (2022). The Bogey Phenomenon in Sport. *IX Mathsport International 2022 Proceedings*. Online/Reading, UK.

Bunker, R., Yeung, C., Susnjak, T., Espie, C., & Fujii, K. (2023). A comparison of the performance of Elo ratings and machine learning techniques for match result prediction in tennis. Paper presented at the *14th European Sport Economics Association (ESEA) Conference* 2023, Cork, Ireland.

Other connected co-authored publications

I was a co-author on the following publications (underlined), which were co-authored with members of the Sports Behaviour Group in the Graduate School of Informatics, Nagoya University. The first publication was accepted for publication at the time of submission of this PhD by prior publication, and the remaining publications were already published.

Fujii, K., Yamada, K., Kono, R., Zhang, Z., & Bunker, R (2025). Machine learning-based analysis of multi-agent trajectories in basketball. To appear as a book chapter in *Artificial intelligence and machine learning in sports science* (Springer).

Scott, A., Uchida, I., Ding, N., Umemoto, R., Bunker, R., Kobayashi, R., ... & Fujii, K. (2024). TeamTrack: A Dataset for Multi-Sport Multi-Object Tracking in Full-pitch Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3357-3366).

Zhang, Z., Bunker, R., Takeda, K., & Fujii, K. (2023). Multi-agent deep-learning based comparative analysis in basketball. In *Proceedings of the 37th National Conference of the Japanese Society for Artificial Intelligence (2023)*, pp. 3U1IS304-3U1IS304. The Japanese Society for Artificial Intelligence, 2023.

Ziyi, Z., Bunker, R., Takeda, K., & Fujii, K. (2023). Multi-agent deep-learning based comparative analysis of team sport trajectories. *IEEE Access*, 11, 43305-43315.

Yeung, C., & Bunker, R. (2023). An events and 360 data-driven approach for extracting team tactics and evaluating performance in football. *Statsbomb Conference 2023 Research Paper*.

Yeung, C., Bunker, R., & Fujii, K. (2023). A framework of interpretable match results prediction in football with FIFA ratings and team formation. *Plos one*, 18(4), e0284318.

Yeung, C., Bunker, R., & Fujii, K. (2024). Unveiling Multi-Agent Strategies: A Data-Driven Approach for Extracting and Evaluating Team Tactics from Football Event and Freeze-Frame Data. *Journal of Robotics and Mechatronics*, 36(3), 603-617.

Yeung, C., Bunker, R., Umemoto, R., & Fujii, K. (2024). Evaluating soccer match prediction models: a deep learning approach and feature optimization for gradient-boosted trees. *Machine Learning*, 1-24.

CHAPTER 1: INTRODUCTION

Technological developments involving video capture and the generation of sports data have markedly increased the availability of such data to athletes, performance analysts, coaches, and management in professional sports. These developments include notational (video analysis) systems, which can be used to annotate events in video and generate event data, as well as optical systems and wearable devices (Ortega & Olmedo, 2017) and computer vision based tracking systems (Thomas, Gade, Moeslund, Carr, & Hilton, 2017; Cioppa et al., 2022; Cui et al., 2024; Scott et al., 2024), which can all be used to generate spatiotemporal tracking data representing the trajectories of players and the ball.

With the effective use of such data, stakeholders in professional sports can make informed decisions, for example, regarding assessing player and team performance and improving player recruitment (Apostolou & Tjortjis, 2019). Given the industry's forecast growth, research that utilises these data sources or proposes new methods to analyse this data is increasingly important. Fortune Business Insights (2024) expects the global sports analytics market, valued at US$3.78 billion in 2023, to expand to US$32.31 billion by 2032.

Determining the factors contributing to successful match outcomes is of fundamental concern to decision-makers in professional sports. Identifying the factors contributing to success has historically depended on coaches' and management's subjective opinions and domain expertise. With technological developments, including notational analysis systems and tracking systems mentioned above, there has been rapid growth in both the volume and variety of data generated in sports, which ostensibly allows for the objective identification of factors contributing to success. However, processing, analysing and extracting useful information from this data can be challenging.

Sports analytics is a discipline that has grown out of fields such as computer science, statistics, and, more recently, data science. On the other hand, sports performance analysis is a field that has its roots in — and is considered a sub-discipline of — sports science (Borms, 2008; Gomez-Ruano, Ibáñez, & Leicht, 2020). Sports performance analysis has traditionally used statistical methods to analyse summarised performance indicator metrics (Hughes & Bartlett, 2002), which are generally computed from event data generated in notational systems or, less commonly, from tracking data (Goes, Kempe, & Lemmink, 2019).

The use of performance indicator metrics that summarise event data generated in notational systems is common in sports performance analysis (Hughes & Bartlett, 2002). However, a more granular analysis of events and augmenting granular event-related information with player movements can allow for a more holistic and in-depth view of performance. With match outcome information, it is possible to determine the importance of particular team-level performance indicators on match results (Robertson, Back, & Bartlett, 2016; Parmar, James, Hughes, Jones, & Hearne, 2017; Young, Luo, Gastin, Tran, & Dwyer, 2019) or player-level (Robertson, Gupta, & McIntosh, 2016). In a univariate manner, these can be identified by analysing the relationship each variable has with the independent variable using statistical methods. This is useful for identifying the most relevant factors contributing to historical match outcomes and areas that can be focused on in future training regimes to improve future performance.

There are, however, some limitations associated with the aggregation of event data into performance indicator metrics (or the aggregation of spatial data into, for example, centroids) in that some granular behavioural and also contextual information is lost (Lames & McGarry, 2007; MacKenzie & Cushion, 2013. Rein & Memmert, 2016). Assessing which are the key (important/relevant) variables in determining outcomes is considered in this thesis using methods that generate interpretable results in the form of decision rules (Bunker & Spencer, 2022) and event sub-sequences (Bunker et al., 2021), and in another study where a boosted tree model with an interpretable structure is utilised (Bunker, Yeung, Susnjak, Espie, & Fujii, 2023).

The two disciplines of sports analytics and sports performance analysis have developed in parallel but largely independently. However, despite differences in methods, terminology, and dissemination outlets, the goals of the performance-focused side of sports analytics and sports performance analysis are mainly similar. The two fields are also becoming more intertwined in recent years, suggesting the emergence of "sports performance analytics". For instance, computer science and data science methods that have traditionally been used in sports analytics, including machine learning and data mining techniques, are being increasingly utilised for sports performance analysis purposes in recent years (for example, in the context of rugby union, see Bennett, Bezodis, Shearer, and Kilduff, 2021; Watson, Hendricks, Stewart, & Durbach, 2020; Bunker et al., 2021).

This thesis contains six interconnected publications, and this contextual statement will demonstrate how these represent a coherent body of work that contributes significantly to knowledge. The publications in this thesis are linked in that all are related to performance in sports, and all publications investigate machine learning in the context of sports for outcome prediction or key pattern identification. There are also methodological, chronological, and scope-related connections between the studies. The connections between publications and how the thesis represents a coherent body of work are discussed in detail in Chapter 3.

In this body of work, generalised conceptual frameworks, surveys, and syntheses are developed that reveal insights in the area of machine learning for sport outcome prediction (Bunker & Thabtah, 2019; Bunker & Susnjak, 2022; Bunker, Yeung, & Fujii, 2025), as well as studies that investigate how machine learning and data mining techniques that have been used in other domains can be transferred to, and be of value in, sporting contexts (Bunker et al., 2021; Bunker & Spencer, 2022; Bunker et al., 2023). A multi-level approach is taken in this thesis in the sense that the publications consider performance at different levels of analysis, for example, at the match outcome, tournament stage outcome, and passage of play levels.

1.1 Research Aims and Contextual Statement Structure

This contextual statement highlights the aims underlying the individual publications and the relationships between the publications contained in this thesis. It will also discuss how combining these papers as a body of work represents a significant and original contribution to knowledge.

Three main research questions are investigated in this thesis:

- (RQ1) How can generalised conceptual frameworks for sports performance analytics be established based on critical analysis and synthesis of existing literature?
- (RQ2) How can data mining methods from other disciplines be best leveraged for key pattern identification in sports performance analytics?
- (RQ3) How can machine learning methods from other domains and contexts be utilised to provide interpretable match outcome prediction in sports performance analytics?

This contextual statement, intended to be read with the enclosed publications, covers six publications in the main text that comprise the PhD by prior publication, along with two

additional publications that appear in the appendix to the thesis (Appendix A2). The two papers in the appendix are connected to the six publications in the thesis. However, they are not explicitly included since they were both accepted after commencing the enrolment period for the PhD by prior publication program.

The remainder of this contextual statement is structured as follows. First, Chapter 2 positions the research conceptually by providing an overview of relevant fields, including sports performance analysis and sports analytics, and also covers relevant background — with a specific focus on machine learning and data mining for match outcome prediction and key pattern identification. Sub-fields of data mining relevant to this thesis's publications, including trajectory mining and sequential pattern mining, are also covered in Chapter 2. Then, Chapter 2 discusses the evolution of sports performance analysis into sports performance analytics and outlines a potential workflow for this emerging area of enquiry. Following this, Chapter 3 describes how the publications in this thesis are linked, demonstrates how the publications together comprise a cluster of original research and highlights the contributions of the body of work. Chapters 4 to 9 contain the six publications that comprise the PhD by prior published work, and each chapter begins with a summary of the publication. Chapter 10 then provides some concluding remarks and discusses potential avenues for further research.

CHAPTER 2: BACKGROUND, CONCEPTUAL POSITIONING & RELATED FIELDS

This chapter aims to position the thesis and its constituent publications within existing disciplines and research areas. It also provides an overview of the fields relevant to the publications comprising this thesis, specifically sports performance analysis, sports analytics, and the emerging area of sports performance analytics. This chapter also covers two of the specific focus areas of this thesis: machine learning and data mining for match outcome prediction and key pattern identification, respectively.

2.1 Sports Performance Analysis

To understand sports performance analysis, it is helpful first to consider what constitutes performance in general. The Oxford Dictionary (s.v. "performance, n., sense 1.b", 2023) defines performance as:

> *"The quality of execution of such an action, operation, or process; the competence or effectiveness of a person or thing in performing an action; spec. the capabilities, productivity, or success of a machine, product, or person when measured against a standard."*

Applying this general definition of performance in the context of sport, performance can be interpreted as the quality of the execution of the actions of the player(s) and their competence and effectiveness in performing actions that contribute to success, where success can be ultimately observed in performance and winning outcomes.

Sports performance analysis grew out of and is now considered a sub-discipline of sports science (Borms, 2008; Gomez-Ruano, Ibáñez, & Leicht, 2020). The field has mainly emerged since the early 2000s. Sports performance analysis can be described as the provision of reliable and valid performance-related information to coaches and athletes with the intention of improving future performance (O'Donoghue, 2014). O'Donoghue (2009) defined sports performance analysis as "the investigation of actual sports performance or performance in training". O'Donoghue noted that what distinguishes sports performance analysis from other areas of sports science is that it is concerned with actual performance (in either matches or training), as opposed to what is observed from laboratory-based activity or qualitative data obtained from self-reports (e.g., questionnaires or interviews).

Hughes & Bartlett (2019) highlighted the crucial role of accurate and objective feedback in the performance improvement process, which occurs between coaches and players in sports, noting that “no change in performance of any kind will take place without feedback”. For this feedback loop between athletes and coaches to develop, there needs to be a means of measuring performance. Notational analysis was proposed as “an objective way of recording performance so that critical events in that performance can be quantified in a consistent and reliable manner” (Hughes & Bartlett, 2019). The utility of notational analysis stems from the fact that even the best coaches only recall 30% to 50% of important events in a match (Franks & Miller, 1991). Notational analysis initially involved the use of pen-and-paper-based systems (Reep & Benjamin, 1968) to record, in a process that is referred to as “annotating”, “tagging”, or “coding”, events that occurred during soccer matches.

As technology has evolved, computer-based notational systems such as Hudl SportsCode (Lincoln, NE, USA) and Dartfish (Fribourg, Switzerland) have made it possible to annotate events from video footage captured in real-time or post-match. The tagged event data can then be exported (e.g., in XML format) from the video analysis system and transformed into performance indicators that summarise historical performance. More specifically, performance indicators are a selection or combination of action variables that aim to define some or all aspects of a sports performance (Hughes & Bartlett, 2002). Performance indicators enable the objective quantification of performance (Vogelbein, Nopp, & Hökelmann, 2014; Parmar, James, Hughes, Jones, & Hearne, 2017) and commonly aggregate or summarise event data, for example, into averages (e.g., the average number of tackles made by a specific player per match), ratios (e.g., the number of shots on target compared to shots attempted), or frequencies (e.g., the total number of penalties conceded by a team during a match). Performance indicators at the team level can be used to analyse current team performance or to track it over time.

In non-team sports such as tennis, a particular player can be compared to other players of similar ability that they compete against. For instance, professional players in the Association of Tennis Professionals (ATP) league can be compared against other players in the ATP tour. In team sports, performance can be analysed at various levels, such as team, player unit/group, and individual player levels. These levels are hierarchical, so performance indicators at the player level can be aggregated to obtain player-, group- or team-level performance indicators.

The performances of two players who play the same position can be compared, for example, by constructing performance profiles (Hughes, Evans, & Wells, 2001; O'Donoghue, 2005; Butterworth, O'Donoghue, & Copley, 2013; O'Donoghue, 2013) based on the players' performance indicators. For instance, to compare the performance of a halfback in rugby union, one could compare their performance profiles with those of the other halfback(s) in the same team and all other halfbacks in other teams in the same competition (or competitions of a similar standard). In sports science, performance profiles were initially constructed by athletes who self-assess their performance and coaches assessing the same performance (Dale & Wrisberg, 1996; Butler, Smith, & Irwin, 1993). However, collecting event data from notational video analysis systems means that performance profiles can now be constructed based on actual performance metrics, giving a more objective view of performance.

The ability for performance analysts to create coding windows and visual user interfaces created within a video analysis system containing buttons to annotate specific events (an example is shown in Figure 1a) means that events of interest to coaches can be tagged in match video and subsequently analysed. By adjusting the coding windows, the captured events can be adjusted over time via the feedback loop between performance analysts and coaches (Figure 1b). Companies such as Stats Perform Opta Data collect event data by tagging events in matches in multiple competitions and sports (Stats Perform a, n.d.). This enables a centralised repository of data for all teams in a competition to be created, which teams can generally access, provided they are willing to pay for it. To provide data that is consistent and ensure that valid comparisons between players and teams can be made, Opta maintains operational event definitions for their coded event data in several sports, such as soccer (Stats Perform b, n.d.), which are standard definitions of what constitutes a specific event in a particular sport (e.g., what constitutes a turnover in soccer). Sometimes these definitions are established by expert consensus; for example, in the case of netball, see Mackay et al. (2023).

a)

b)

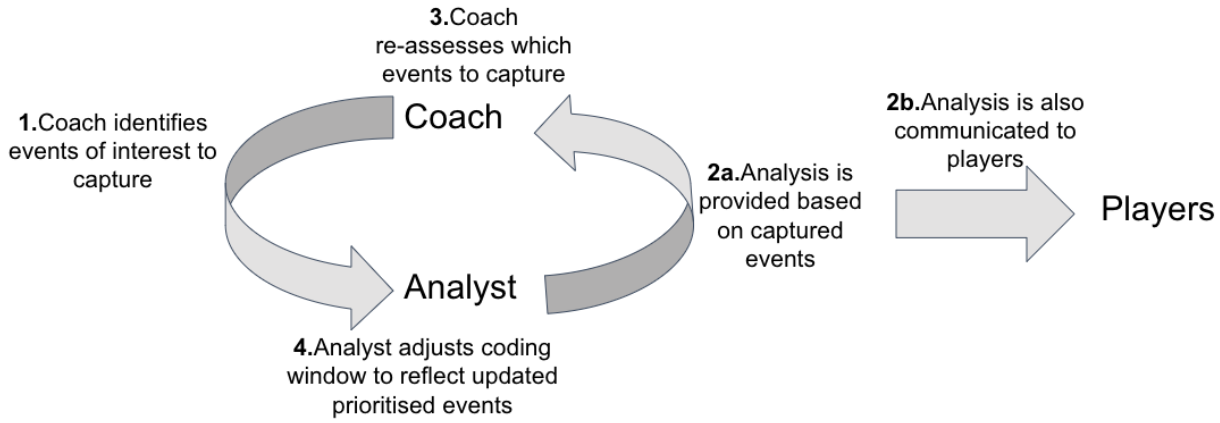


Figure 1: Coding windows and feedback loops. *a) An example of a coding window for Touch Rugby in Hudl SportsCode, with buttons that can be used to annotate (tag/code) specific events and can be adjusted as events prioritised by coaches change over time. b) Illustration of the feedback loop that exists between coaches and performance analysts. Source: https://support.hudl.com/s/article/code-window-modes-sportscode?language=en_US*

Research in sports performance analysis has primarily focused on identifying key performance indicators relevant to success by analysing the differences in game actions between successful and less successful teams (Lord, Pyne, Welvaert, & Mara, 2020). Some studies have considered performance indicators contributing to successful and unsuccessful outcomes, which often aim to identify key performance indicators that statistically significantly discriminate between successful and unsuccessful outcomes, for example,

winning and losing (Hughes, Barnes, Churchill, & Stone, 2017; Lago-Peñas, Lago-Ballesteros & Rey, 2011; García, Ibáñez, De Santos, Leite, & Sampaio, 2013; Bunker & Spencer, 2022).

However, Lames and McGarry (2007) suggested that some performance indicators may not be consistently accurate or provide valuable player or team performance insights and that performance analysis may need to adequately account for the complexity of a sport, its dynamic interactions, and sequential nature. The authors emphasised the need for more sophisticated performance indicators incorporating contextual information. The use of performance indicators in isolation, without accounting for confounding factors or context, has also been critiqued by MacKenzie and Cushion (2013). Rein & Memmert (2016) further echoed these sentiments, noting that one of the main limitations of notational systems and analysis of performance indicators is that nearly all contextual information is discarded.

Contextual (situational) information can include factors such as the stage of the season, match location, and opposition quality. Contextual variables are often external to matches; that is, they are not related to events that occur within matches. It can be essential to distinguish contextual factors from performance indicators since these are generally out of a team's control. For instance, teams cannot change the season's stage, the match's location, or the quality of an opposition team, regardless of improvements in their team performance.

In light of these limitations, researchers have increasingly used dynamic and complex analyses for a deeper understanding of performance in sports (O'Donoghue, 2009, as cited in Gomez-Ruano, Ibáñez, & Leicht, 2020). For instance, multivariate and holistic approaches for understanding sports performance that combine variables from, for example, physical, technical, and tactical contexts are becoming increasingly important (Memmert, Lemmink & Sampaio, 2017; Rein & Memmert, 2016). More multivariate, dynamic, complex, and holistic approaches can be achieved by augmenting the concepts of sports performance analysis with sports analytics, which is discussed in the following subsection.

2.2 Sports Analytics

In contrast to sports performance analysis, which primarily developed out of sports science, sports analytics has its roots mainly in fields such as statistics (and, more recently, data science) and computer science (Figure 2). Sports analytics emerged as a distinct discipline in the 1990s (O'Donoghue, 2014) and has expanded rapidly in terms of research interest in the 21st century. The discipline partly grew from the interest in the "Moneyball" story, originally a book by Michael Lewis (Lewis, 2004) that was later made into a Hollywood film of the same name. "Moneyball" describes the story of the manager of the Oakland Athletics Major League Baseball team, Billy Beane, who used methods based on analytics to identify low-salary but potentially high-value players to scout, significantly improving team performance. Interest in the domain has expanded since the 2010s, with several sports analytics textbooks (e.g., Miller, 2015; Alamar, 2013; Severini, 2014; Jayal, McRobert, Oatley, & O'Donoghue, 2018) having been published during that decade.

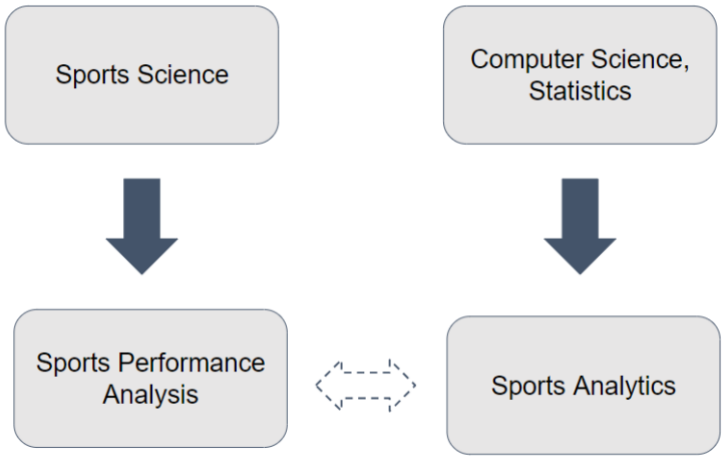


Figure 2: Origins of sports performance analysis and sports analytics. *Sports performance analysis has its roots in sports science, while sports analytics comes primarily from computer science, statistics, and, more recently, data science. However, as the dashed double-headed arrow indicates, the boundary between the two fields is becoming blurry over time.*

Sports analytics has been defined as (Alamar & Mehrotra, 2011 as cited in Alamar, 2013):

> "*the management of structured historical data, the application of predictive analytic models that utilize that data, and the use of information systems to inform decision*

makers and enable them to help their organizations in gaining a competitive advantage on the field of play."

Jayal et al. (2018) point out that the above definition of Alamar and Mehrotra (2011) contains three distinct steps: data management, modelling, and the use of information systems. These three steps are depicted in Figure 3.

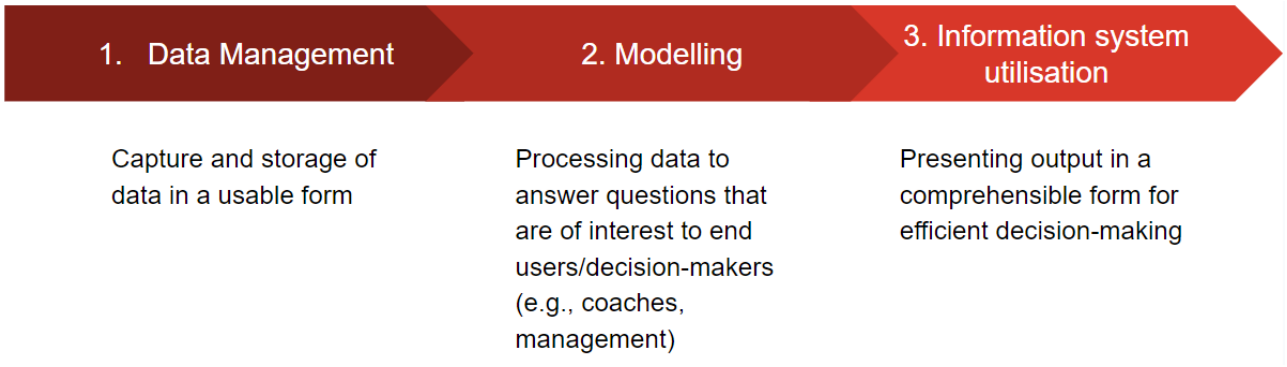


Figure 3: Sports Analytics Process. *The three steps comprising the sports analytics process, as identified by Jayal et al. (2018).*

Although the first step of the process, data management, is certainly an important consideration from a practical perspective[1], studies in this thesis focus more on steps 2 and 3.

Sports analytics has numerous applications, including predicting the performance of players and teams, estimating the market value of players, and predicting the occurrence of injuries (Apostolou & Tjortjis, 2019). Given that performance is considered in both sports analytics and sports performance analysis, sometimes the two fields are referred to interchangeably in the literature. For instance, Morgulev & Lebed (2024), while recognising the definition of Alamar & Mehrotra (2011) and Alamar (2013) above, used the term sports analytics in a more narrow form by referring to "sports analytics" and "performance analysis" interchangeably throughout their paper. On the other hand, to distinguish between the performance-related and economic aspects, the terms "sports performance analytics" and "sports management analytics" are sometimes used in American terminology (Link, 2018).

[1] For example, a practical consideration in Bunker, Le Duy, Tabei, Takeuchi, & Fujii (2024), which is in the appendix to this thesis, was the use of a PostgreSQL database to house spatiotemporal data and geospatial libraries in Python, such as geopandas to interact with and process this data. However, the paper's main contribution was proposing a novel algorithm to comprehend this data.

The current thesis's studies are related to the performance aspect of sports analytics rather than management or economic aspects.

The boundary between sports management analytics and sports performance analytics is sometimes not clearly demarcated. For instance, the estimation of player value can relate to player value in a monetary sense in terms of their market value if they were to be transferred or their value in terms of contribution to performance, for example, considering the value of a player's on-ball actions (Decroos, Bransen, Van Haaren, & Davis, 2020). There is an apparent link between the on-field performance of a player and their market value; for example, see Behravan & Razavi (2021) and Berri, Butler, Rossi, Simmons, & Tordoff (2024) for the case of determining the transfer market values of goalkeepers in professional soccer. Similarly, attendance forecasting can also require considering team-level performance since a team performing better will generally attract larger audiences to their matches. Nonetheless, attendance forecasting is usually considered in sports management or sports economics. In addition, it is generally clear whether a study has primarily a sports management focus or mainly a sports performance focus based on the study's research goals.

The remainder of this section will discuss background in relation to specific areas of sports analytics, particularly machine learning and data mining, with a focus on sports match outcome prediction and key pattern identification.

2.2.1 Machine Learning

2.2.1.1 Machine Learning and its applications in sports

Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that focuses on the use of data and algorithms to enable AI to imitate the way that humans learn (IBM, 2024). ML involves the application of algorithms or models to learn patterns and relationships in data. The two main types of machine learning are supervised machine learning and unsupervised machine learning, which involve taking as input labelled and unlabelled data, respectively. Supervised machine learning can be further subdivided into classification and regression, which involve predicting a discrete or numeric target variable, respectively (Figure 4).

Machine learning has found numerous applications in sports analytics, for example, in injury prediction (Rommers et al., 2020; Van Eetvelde, Mendonça, Ley, Seil, & Tischer, 2021),

movement recognition (Cust, Sweeting, Ball, & Robertson, 2019; Giles, Peeling, Kovalchik, & Reid, 2023), sports betting[2] (Hubáček, Šourek, & Železný, 2019a), training (Me & Unold, 2011), and for predicting the performance of players (Pappalardo et al., 2019; Zhu & Sun, 2020) or teams. Player-level performance indicators can be grouped into single-action and all-action indicators (Davis et al., 2024), such as expected goals (xG) (Green, 2012) and the Valuing Actions by Estimating Probabilities (VAEP) method (Decroos, Bransen, Van Haaren, & Davis, 2019), respectively. In particular, xG considers a single action, shots, analysing the probability of success of a particular shot in soccer by training a machine learning model based on historical shot characteristics. In contrast, VAEP considers all on-ball actions (e.g., pass, cross, shot) made by a player and assigns a value to each action.

The adoption of machine learning in the context of sports can be challenging. Coaches are often former players and have deep domain knowledge of what leads to successful performance in a particular sport. Thus, there is a need to emphasise that the role of machine learning is to act as a decision support tool to augment their expertise and to uncover insights that may not have been immediately obvious, not to replace their expertise. The effectiveness of implementing machine learning techniques in practice also depends on the data available, and having the digital infrastructure in place to capture necessary data, for example, event data via video analysis systems, and movement data through wearable devices, optical systems, or computer vision systems. Machine learning is currently largely employed in specific professional sports; soccer and basketball have historically been leaders in its utilisation. While statistical methods have traditionally been taught in sports performance analysis courses, an increasing focus on advanced analytics techniques including machine learning will help increase its adoption in practice. A focus on data visualisation and communication in such courses will also help ensure performance analysts are able to effectively deliver insights obtained from these techniques and to demonstrate clear benefits from their use.

Another application of machine learning is for predicting sports match results, which will be discussed in the next subsection.

[2] As well as maximising predictive performance to compete in competitions, another common use of machine learning match outcome prediction models is for sports betting (Hubáček, Šourek, & Železný, 2019a; Stübinger, Mangold, & Knoll, 2019; Wilkens, 2021). However, this topic lies outside of the scope of this thesis.

2.2.1.2 Machine Learning for sports match outcome prediction

While statistical models were traditionally used for match outcome prediction in sports (Dixon & Coles, 1997; Maher, 1982), machine learning has been increasingly used over the past two decades (Bunker & Susnjak, 2022). Constantinou (2019) categorised soccer result prediction models into three groups: statistical models, machine learning and probabilistic graphical models, and rating systems. However, Bunker, Yeung, & Fujii (2025) pointed out that it may not make sense to split ratings and machine learning into two groups since some top-performing recent studies (e.g., Berrar, Lopes, & Dubitzky, 2019; Hubáček, Šourek, & Železný, 2019b; Razali, Mustapha, Mostafa, & Gunasekaran, 2022) used rating systems as features in machine learning models. Rating systems can, therefore, be used as a predictive model in their own right by computing the rating and predicting the team with the higher rating in a particular match as the winner, or ratings can be used as features in machine learning models. Although one of the first studies to use machine learning to predict sports outcomes (Purucker, 1996) employed both supervised learning (artificial neural networks - ANNs) and unsupervised learning (clustering), the far more common approach for sports match result prediction using machine learning in subsequent studies has been supervised learning.

Classification models for match result prediction generally involve predicting a win/draw/loss target variable in sports with three possible outcomes (e.g., soccer) or predicting a win/loss target variable in a sport with two possible outcomes (e.g., basketball, tennis). Conversely, regression involves predicting a numeric target variable such as the score or score margin (Figure 4) — the difference in points/goals scored between the opposing teams or players. Classification models have been more widely employed than numeric prediction models in the literature. However, some authors have compared the performance of the two types of models (e.g., Delen, Cogdell, & Kasap, 2012 found classification models to be superior to numeric prediction in their study that considered a dataset consisting of American Football matches). In addition, the recent 2023 Soccer Prediction Challenge, the leading approaches from which were published in an issue of the Machine Learning (Springer) journal (one of which was Yeung, Bunker, Umemoto, & Fujii, 2024), involved participants developing models for two tasks. The first task involved generating probabilities for win/draw/loss outcomes for each match, and the second involved predicting the exact goals scored by each team (from which the match result can then be derived). Although the rankings of team names of participants in the 2023 competition are available online, the full details of models developed by the teams are still not fully available at the time of writing. However, the full

results of the 2017 iteration of the competition are available and are discussed, along with some subsequent studies that have used the same dataset, the Open International Soccer Database, in subsections 2.2.1 and 2.2.2 in Chapter 9 of this thesis. Gradient boosted tree models applied to ratings were found to be strong-performing models in the competition and have proven successful in subsequent studies using the same dataset. Some established machine learning models, including random forests and ANNs, as well as high-performing contemporary gradient-boosted tree models such as Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016) and Categorical Boosting (CatBoost) (Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2018) can be utilised for both regression and classification tasks.

The performance of players or teams is directly related to match outcome prediction since the result of a match is heavily dependent on the performances of all players in a team in team sports (or on the performance of the player in a solo or dual sport[3]). As mentioned, however, situational (contextual) factors such as stage of the season, game location, and opposition quality can also affect match outcomes but are not controllable by teams, so they are important in controlling for the performance of teams in competitions (Gomez, Lago-Peñas, & Sampaio, 2013). It is for this reason that modern rating systems in soccer, for example, Berrar ratings (Berrar, Lopes, & Dubitzky, 2019), Generalised Attacking Performance (GAP) ratings (Wheatcroft, 2020), and pi-ratings (Constantinou & Fenton, 2013) often build separate home and away ratings for teams to control for match location, and also often construct separate offensive and defensive ratings in order to generalise the attacking and defensive capability of teams across all teams in the competition(s).

[3] A dual sport is played between opposing players and includes sports such as tennis, badminton, fencing, etc. (Palut & Zanone, 2005)

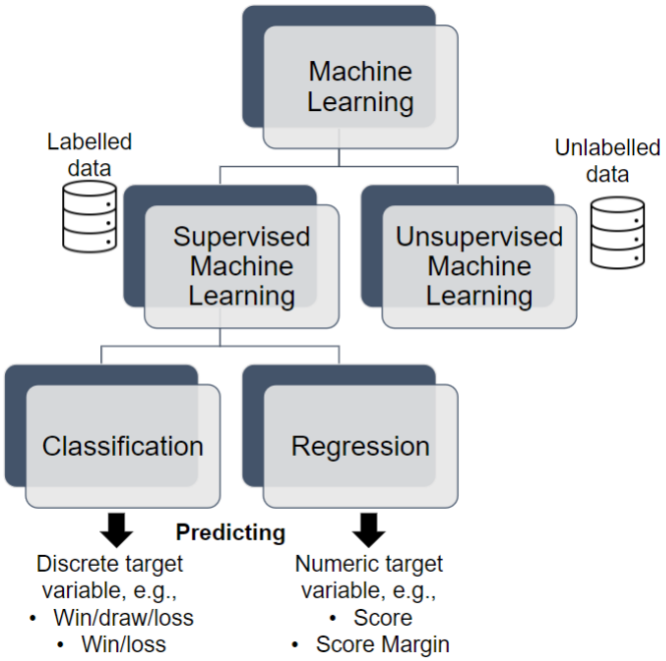


Figure 4: Types of machine learning. *Machine learning models can be categorised into two main types: supervised and unsupervised, which are applied to labelled and unlabeled data, respectively. Supervised learning can be further subdivided into classification and regression, which involve predicting a discrete or numeric target variable.*

2.2.1.3 Interpretability of models and results

In the literature, machine learning models predicting match outcomes have often not provided meaningful output (Elstak, Salmon, & McLean, 2024). One reason for this is that the main focus of the literature has prioritised predictive performance over model interpretability. Match outcome prediction models that are interpretable can be helpful for prediction and coaches and performance analysts in identifying the most pertinent variables that contribute to winning.

The interpretability of the alternating decision tree (Freund & Mason, 1999), which utilises the boosting algorithm AdaBoost (Freund & Schapire, 1997) to grow the tree but has an interpretable tree structure, made it appealing to use as one of the candidate models to predict match results in tennis and compare with ratings-based methods in Bunker et al., (2023). As discussed later, other studies have used interpretable methods such as chi-squared automatic interaction detection (CHAID) decision trees (Robertson, Back, & Bartlett, 2016; Parmar et al., 2017) for this purpose.

More recently, to increase interpretability and determine the effects of specific features on the target variable, studies in sporting contexts have employed explainable AI (xAI) techniques such as TreeExplainer SHapley Additive exPlanations (SHAP) (Lundberg, 2017) (e.g., Moustakidis, Plakias, Kokkotis, Tsatalas, & Tsaopoulos, 2023; Geurkink et al., 2021; Ren & Susnjak, 2022), random forest feature importance (Eryarsoy & Delen, 2019; Groll, Ley, Schauberger, Van Eetvelde, 2019), and aggregated profiles (Cavus & Biecek, 2022).

Other approaches to incorporating interpretability include investigating, in a univariate fashion, the relationships between each model feature and the target variable (as mentioned in the introduction) or, as in two of the publications in this thesis, utilising models that generate interpretable results in the form of decision rules (Bunker & Spencer, 2022) or event patterns (Bunker et al., 2021).

2.2.1.4 Common ML model evaluation metrics for match result prediction

As in other domains, machine learning models for sports result prediction aim to optimise performance in terms of a specific measure. For example, accuracy has been a standard performance metric for classification models in sports match outcome prediction in many studies. Some measures are more appropriate for imbalanced data, for example, balanced accuracy, the F1-score, and the area under the precision-recall curve. However, sports match outcome data with home win and away win tends to only exhibit slight imbalance due to the home advantage effect (Schwartz & Barsky, 1977 as cited in Ryall & Bedford, 2010). Thus, accuracy is often sufficient as an evaluation metric in sports with two outcomes in many cases. Accuracy is given by simply dividing all correct predictions by the total number of predictions made or using the values from a 2-by-2 confusion matrix:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Ranked probability score (RPS) (Constantinou & Fenton, 2013; Epstein, 1969; Murphy, 1969; Murphy, 1970) is the current standard for models that output probabilities (especially for three-class problems to output the probabilities of a win, loss, and draw). The advantage of RPS is that it considers the ordinal structure of the classes. For instance, if a model predicts a win for a particular team but the result is a draw, this is a more accurate prediction than if the result was a loss. RPS is given by:

$$\text{RPS} = \frac{1}{r-1}\sum_{i=1}^{r-1}\left(\sum_{j=1}^{i}(p_j - a_j)\right)^2 \quad (2)$$

Where *r* denotes the number of potential match outcomes, *p* is the probability of a particular outcome predicted by the model, and a is a binary variable indicating the actual match result. The RPS values always lie within the 0 and 1 (inclusive) range, with a lower RPS indicating a better prediction. In particular, an RPS value of 0 indicates a perfect prediction by a model, while a value of 1 represents an entirely incorrect prediction. Although RPS has become commonly used in recent studies for models that output probabilities, the ignorance score (Good, 1992; Roulston & Smith, 1992), which involves taking the logarithm of the outcome probability, has been suggested as a metric for model evaluation in this domain that has preferable properties to RPS (Wheatcroft, 2021).

Root mean squared error (RMSE) is the standard metric for numeric prediction in this domain, and is given by:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (3)$$

where N denotes the total number of data points, y_i denotes the actual observed values, and $\hat{y_i}$ represents the values predicted by the model. Models that yield a lower RMSE are preferred.

For the win/draw/loss outcome probability prediction task in the 2023 Soccer Prediction Challenge, submitted models were evaluated based on the RPS. In the second task, RMSE was used as the evaluation metric to predict the exact goals scored by each team. Evaluation metrics for sports outcome prediction machine learning models are covered in more detail in Bunker, Yeung, and Fujii (2025).

2.2.1.5 Incorporation of domain knowledge & limitations of benchmark datasets

There remains a gap between machine learning researchers who develop match outcome prediction models and the sports performance analysis research community (Bunker & Susnjak, 2022). The latter generally have better domain expertise and knowledge of the factors that drive successful performance in a particular sport. As with machine learning in general, the incorporation of sporting domain knowledge into the machine learning modelling

process is vital for enhancing predictive performance (Joseph, Fenton, & Neil, 2006; Berrar, Lopes, & Dubitzky, 2019; Berrar, Lopes, & Dubitzky, 2024).

Benchmark-type datasets for sports match outcome prediction such as the Open International Soccer Database (Dubitzky, Lopes, Davis, & Berrar, 2019), used in the 2017 Soccer Prediction Challenge, and the 2023 Soccer Prediction Challenge dataset only contain goals scored as the features related to in-match events. Even if the goals are condensed into pi ratings or Berrar ratings, the models are not especially useful from a performance analysis perspective since it is trivial that better offensive performance and scoring more goals, and better defensive performance and conceding fewer goals result in better performance.

The increased availability of datasets in the public domain that contain varied types of data encountered in sports, such as spatiotemporal, event, and situational datasets, could help to advance further research innovation from the machine learning research community. Given access to such data is often challenging for commercial reasons, releases of datasets with sensitive variables anonymised or encrypted could still be valuable. The generation of synthetic datasets, as has been explored in domains such as healthcare where data access is tightly controlled for privacy reasons (Gonzales, Guruswamy, & Smith, 2023), is another possible approach.

Along with traditional statistical methods taught in performance analysis courses at the university level, machine learning may also be useful knowledge for aspiring sports performance analysts to aquire, and the integration of machine learning into curricula and courses could help develop technical skills within teams and drive collaborative innovation. Mateus et al. (2024) suggests academic institutions adjust their curricula to provide sports science students with skills in data analysis and visualisation, machine learning, ethical considerations, and the translation of results into practice. Forums such as conferences that involve researchers and practitioners from both sports performance analysis, and researchers from the machine learning community, will also be valuable in addressing the gap between the two communities.

A key area in which performance analysts and coaches can assist in advancing machine learning research is in the selection and engineering of features. Domain expertise is important in terms of identifying potential pertinent features to be tested, potentially against

or combined with feature selection algorithms. This could lead to increases in predictive accuracy, and the involvement of sporting experts in the machine learning research and processes will lead to increasing familiarity with the techniques and perhaps a greater willingness to make use of such machine learning in practice. The involvement of sports performance practitioners and/or researchers in machine learning studies will also assist in the interpretation of results, which will, in the case of black-box models, require the use of explainable AI methods such as SHAP. Both Chmait & Westerbeek (2021) and Mateus et al. (2024) point out that AI-based systems including those built using machine learning will not replace human experts in sports, but will rather complement and augment their specialist knowledge for use as decision support tools to improve outcomes. Therefore, those who embrace such tools, rather than view them as a potential threat, will be able to gain a competitive advantage.

2.2.1.6 Machine Learning in this thesis

As has been mentioned, one of the ways in which the six studies in this thesis are linked is that all investigated machine learning techniques. Some of the publications focused on machine learning for sports match outcome prediction. For instance, Bunker and Thabtah (2019) proposed a practically implementable conceptual framework for sports match outcome prediction using machine learning after critically analysing the literature related to the use of ANNs for sports match result prediction. While Bunker & Thabtah (2019) covered both individual and team sports, Bunker & Susnjak (2022) focused on team sports specifically, providing a critical analysis and synthesis of studies conducted between 1996 and 2019 in match result prediction in team sports, including invasion sports[4] and striking/fielding sports. Bunker, Yeung, and Fujii (2025) subsequently covered machine learning for match outcome prediction in a specific invasion sport, soccer, and more in-depth coverage of the topic, including state-of-the-art approaches and potential future research avenues, could be provided as a result. Bunker et al. (2023) applied the framework proposed in Bunker & Thabtah (2019) in practice, investigating match outcome prediction in a dual sport, tennis. The study conducted a comparative evaluation of ratings-based methods: Elo ratings and an extension of Elo, Weighted Elo (WElo) (Angelini, Cantila, & De Angelis, 2022)

[4] Invasion sports are team sports in which the main goal is to invade the opposition territory and score more goals than the opponent within an allocated time frame (Webb, Pearson, & Forrest, 2006). Thus, invasion sports are said to be time dependent. Common invasion sports include soccer, hockey, basketball, and rugby. In contrast, striking/fielding sports such as baseball and cricket do not have allocated time periods but, rather, are innings dependent (see Figure 12 in Appendix A1, or Read & Edwards, 1992, as cited in Hughes & Bartlett, 2002 for further classifications of invasion sports and striking and fielding sports).

that places higher weight on the most recent match result against a set of machine learning techniques including an interpretable boosted tree model (ADTrees) (Freund & Mason, 1999).

Other studies in the thesis focused on using machine learning for key pattern identification. Bunker et al. (2021) leveraged a discriminative sequential pattern mining technique that is also a machine learning classifier (S3P-classifier) (Nakagawa, Suzumura, Karasuyama, Tsuda, & Takeuchi, 2016; Sakuma et al., 2019) to identify key patterns in the form of event subsequences that discriminated between scoring and non-scoring outcomes in passages of play in rugby union. Bunker & Spencer (2022) compared the utility of an interpretable rules-based machine learning algorithm, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) (Cohen, 1995), alongside the traditionally utilised statistical methods (Wilcoxon signed rank test and Cohen's d effect sizes - Cohen, 1988). RIPPER identified key patterns in the form of performance indicator variables or combinations of these variables, as well as their value ranges, which were relevant for success at the play-off and group stages of the 2019 Rugby World Cup (RWC).

2.2.2 Data Mining

Data mining incorporates techniques such as clustering, classification, outlier analysis, and pattern mining and involves extracting understandable information from databases for decision-making (Fournier-Viger, Lin, Kiran, Koh, & Thomas, 2017). Along with machine learning, data mining forms another set of techniques applied in sports analytics. There are overlaps between machine learning and data mining, and they are sometimes considered to be within the same discipline. Machine learning algorithms are sometimes considered to be a data mining technique. For instance, in the Knowledge Discovery in Databases (KDD) process proposed by Fayyad, Piatetsky-Shapiro, & Smyth (1996), a framework to guide the application of data mining techniques in practical settings, data mining is one step in the process which could involve applying potentially any knowledge discovery algorithm or model. Furthermore, the inclusion of both "data mining" and "machine learning" in the title of the foundational machine learning textbook by Witten, Hall, & Frank (2011), entitled "Data Mining: Practical Machine Learning Tools and Techniques", also suggests a strong link between data mining and machine learning (note the inclusion of classification, a type of supervised learning, in this definition — again highlighting the vital link between the two fields).

Data mining and machine learning algorithms also have commonality as knowledge discovery techniques in that they differ from traditional statistical methods in terms of their epistemology, that is, the a priori knowledge that the modeller or machine is assumed to possess. In contrast to statistical inference, which requires defined (null and alternative) hypotheses, predefined hypotheses are not generally specified when applying data mining and machine learning methods. Instead, as the term "knowledge discovery" implies, knowledge is derived from patterns recognised in the data as the methods are applied, and this discovered knowledge is analysed after the methods have been applied. According to Mohaghegh (2021), traditional statistical inference involves a deductive approach to knowledge, whereas machine learning and data mining take an inductive approach. As Goes, Brink, Elferink-Gemser, Kempe, and Lemmink (2021) put it; data mining involves identifying robust patterns from data without the need to formulate hypotheses about their existence.

The robustness of data mining methods can be evaluated by, for instance, computing pattern support, which is the frequency with which a particular pattern appears in a dataset. Whether patterns are significant and unlikely to be found by chance can be identified using techniques like subgroup discovery (Grosskreutz & Rüping, 2009). Alternatively, in sports, a decision maker, such as a coach, can subjectively rank relevant patterns in order of importance, and patterns can then be weighted using a weighting function (Decroos, Van Haaren, & Davis, 2018). Alternatively, as Bunker et al. (2021) proposed, discriminative sequential pattern mining can be employed to objectively identify relevant patterns in subsequences from passage of play event sequences.

The remainder of this subsection discusses two specific types of data mining relevant to this thesis: trajectory mining, which is commonly applied to spatiotemporal tracking data, and sequential pattern mining, which can be applied to event data.

2.2.2.1 Trajectory mining & spatio-temporal tracking data

Trajectory mining is a sub-discipline of data mining (Figure 5). A data source increasingly used with machine learning and data mining in sports is spatio-temporal tracking data (Gudmundsson & Horton, 2017; Kovalchik, 2023). The volume of spatio-temporal data generated in sports has expanded dramatically due to the development of player tracking systems in the form of wearable technology devices, computer vision, and optical systems

(Scott et al., 2024). Historically, tracking data were mainly used by strength and conditioning specialists to monitor athlete demands and loads (Sarmento et al., 2014), for example, to determine when to replace players during a match in order to optimise performance and reduce the likelihood of injuries occurring. However, more recently, spatio-temporal tracking data has also begun to be used to analyse athlete movements and behaviour.

Szymanski (2020) claimed that sports analytics research requires a greater emphasis on solid theoretical foundations to make meaningful sense of collected data. In the context of spatio-temporal tracking data specifically, Goes et al. (2021) suggested a greater degree of two-way collaboration between computer scientists and sports scientists. The authors noted that computer scientists have the technical ability to develop novel computational methods to process and analyse spatio-temporal tracking data. In contrast, sports scientists generally have better sporting domain knowledge. The review paper of Rein & Memmert (2016) also identified the opportunity for increased collaboration between the computer science and sports science research communities in the context of utilising spatio-temporal tracking data in soccer. Rein and Memmert noted that most machine learning-related studies in sports analytics that have leveraged tracking data have been carried out by researchers from the computer science field with little apparent involvement from sports science researchers.

It has been noted that data management systems that handle spatiotemporal tracking data in soccer have received greater attention in the literature than making sense of this data, for instance, through its aggregation into interpretable features that capture the complex dynamics of tactical behaviour (Goes et al., 2021). Behaviour has been analysed by summarising spatio-temporal data into aggregated features, for example, centroids, surface areas, or Voronoi diagrams (Frencken, Lemmink, Delleman & Visscher, 2011; Memmert, Lemmink & Sampaio, 2017). However, this may lead to relevant complex tactical behaviour being overlooked. To resolve this, composite spatial aggregation has been proposed in soccer. For instance, Goes, Kempe, Meerhoff, and Lemmink (2019) constructed a pass effectiveness measure in soccer using a composite measure of aggregated features that incorporated centroids, team spread, and surface area. Spatial aggregation reduces system complexity to an interpretable level by deriving features that capture group-level behaviour within specific time intervals (Goes et al., 2021). These may be fixed intervals based on a specified period (e.g., a half, quarter, or minute) or an event-based interval considering the period before or after a specific event. An advantage of data mining techniques is that they

can often be applied directly to tracking data without needing to carry out aggregation (Goes et al., 2021).[5]

Advances in object location measurement technology mean discovering knowledge from trajectory data through trajectory mining is becoming increasingly important (Zheng, 2015, as cited in Le Duy et al., 2020). The input data for trajectory classification algorithms are sequences of spatio-temporal points (Ferrero, Alvares, Zalewski, & Bogorny, 2018). In trajectory mining, a relevant sub-trajectory is the portion of a trajectory that can potentially discriminate between classes of the classification problem (Ferrero et al., 2018). Ferrero et al. (2018) proposed a method to identify relevant sub-trajectories based on Shapelet (Ye & Keogh, 2011) analysis but adapted it to trajectories rather than time series.

Spatiotemporal tracking data has also been used to automatically detect formations and playing styles in sports. For instance, Bialkowski et al. (2014) proposed a role representation approach to assign consistent roles to soccer players, enabling the comparison of formations across different teams and matches and, in turn, the detection of team styles. For tactical pattern detection, other studies have used clustering techniques such as K-means and Ward hierarchical clustering on flow motifs constructed from passing patterns (Gyarmati, Kwak, & Rodriguez, 2014; Bekkers & Dabadghao, 2019) and dimensionality reduction approaches, including self-organising maps (Grunz, Memmert, & Perl, 2012).

Discriminative sub-trajectory mining aims to identify sub-trajectories that are more similar to sub-trajectories in one group and less similar to sub-trajectories in the other group (Le Duy et al., 2020). Discriminative sub-trajectory mining can be considered a sub-discipline of sub-trajectory mining, which is a sub-discipline of trajectory mining, which is, in turn, a sub-discipline of data mining (Figure 5). Bunker et al. (2024), presented in Appendix A2, proposed a novel multi-agent statistically discriminative sub-trajectory mining algorithm (MA-Stat-DSM) and demonstrated its utility using spatiotemporal tracking data from publicly available SportVU NBA Basketball player trajectory data from the 2015/16 season. The MA-Stat-DSM algorithm extended the Statistically Discriminative Sub-trajectory Mining (Stat-DSM) method of Le Duy et al. (2020) to allow for the trajectories of multiple agents, for

[5] Indeed, the multi-agent statistically discriminative sub-trajectory mining method proposed by Bunker et al. (2024), which is provided in Appendix A2 in this thesis, also does not require spatial aggregation. The algorithm also does not require temporal aggregation, although for interpretability we considered an event-based interval from when the last passer receives the ball until a shot attempt is made or a turnover occurs.

instance, players from each opposing team and the ball. The method automatically identified the most important sub-trajectories that rendered a passage of play effective or ineffective.

Therefore, both Bunker et al. (2024) and Bunker et al. (2021) considered performance at the passage of play level. However, the studies differed in the type of data used. The former study primarily considered tracking data (the time between coordinates was fixed, so only the spatial aspect of the data needed to be considered), although some events, such as passes and shots, could be inferred by analysing the movement of the players and the ball. On the other hand, Bunker et al. (2021), which is included in the main text of this thesis and will be discussed further in the following subsection, utilised event sequences derived from event log data.

Figure 5: Sub-disciplines of trajectory mining*. Hierarchy showing trajectory mining sub-disciplines. In the appendix to this thesis, Bunker et al. (2024) proposed a multi-agent statistically discriminative sub-trajectory mining method.*

2.2.2.2 Sequential Pattern Mining

Pattern mining involves discovering relevant patterns in databases without considering sequence order, while sequential pattern mining also aims to identify meaningful patterns but also accounts for the order of sequences (Fournier-Viger et al., 2017). This makes sequential pattern mining a suitable approach to analyse ordered sequences of events that arise in sports (especially invasion sports). Sequential pattern mining (Agrawal & Srikant, 1995) is also considered a sub-discipline of data mining (Figure 6). As mentioned previously, event log data is generated in sports performance analysis by annotating events in

notational systems. Performance indicators can be computed by transforming these event logs for analysis. However, similar to when spatio-temporal data is aggregated into, for instance, centroids, there is a loss of information associated with aggregating event log data into performance indicators. An approach to analyse event data is to (also) convert it into sequences of events and to leverage sequential pattern mining to gain additional insight into player and team behaviour at a more granular level.

Initial work on the automatic identification of patterns in sports such as soccer included the development of temporal patterns (T-patterns) (Magnusson, 2002; Camerino, Chaverri, Anguera, & Jonsson, 2012). Some studies have identified frequent patterns of movement using spatiotemporal data (Sweeting, Aughey, Cormack, & Morgan, 2017; White, Palczewska, Weaving, Collins, & Jones, 2021). In the context of sequential pattern mining, unsupervised sequential pattern mining algorithms had previously been applied to unlabelled sequence data in sports. For instance, the Co-occurrence Map Sequential Pattern Analysis Method (CM-SPAM) (Fournier-Viger et al., 2014) had previously been applied to the tactical analysis of judo (La Puma & de Castro Giorno, 2017). Sequential Pattern Discovery using Equivalent Class (SPADE) (Zaki, 2001) has been applied to identify frequent and interesting patterns in cycling (Hrovat, Fister Jr, Yermak, Stiglic, & Fister, 2015). An extension of SPADE, CM-SPADE (Fournier-Viger et al., 2014) had been applied to identify frequent sequential patterns in soccer, and the obtained patterns could be weighted based on their relevance according to an end user (e.g., coach) (Decroos, Van Haaren, & Davis, 2018). The studies mentioned above generate pattern support, which is a valuable measure but only indicates how frequently a pattern appears in a dataset, not how relevant the patterns are to specific outcomes (i.e., whether the patterns are associated with successful or unsuccessful outcomes). As mentioned, Decroos, Van Haaren, and Davis (2018) addressed this by introducing a ranking function that enabled a domain expert to place a higher weight on features deemed of greater relevance by this user. However, the limitation of this approach is that it relies on the subjective judgement of the user as to which features are of most significant importance.

As depicted in Figure 6, discriminative sequential pattern mining is a sub-discipline of sequential pattern mining, which is, in turn, a sub-discipline of pattern mining, which is, in turn, a sub-discipline of data mining. Discriminative sequential pattern mining is also known as supervised sequential pattern mining since it is applied to labelled data. The S3P-classifier technique (Nakagawa et al., 2016; Sakuma et al., 2019) utilised by Bunker et al.

(2021) is a classification model incorporating machine learning and discriminative sequential pattern mining. The advantage of the S3P-classifier is that it can be applied to labelled event sequences representing passages of play that are labelled based on whether a team scored in that passage. Key patterns in the form of event subsequences that discriminate between scoring and non-scoring outcomes can be determined using these labelled sequences. This method enables the objective identification of the most important patterns based on the degree to which they discriminate between the outcome labels.

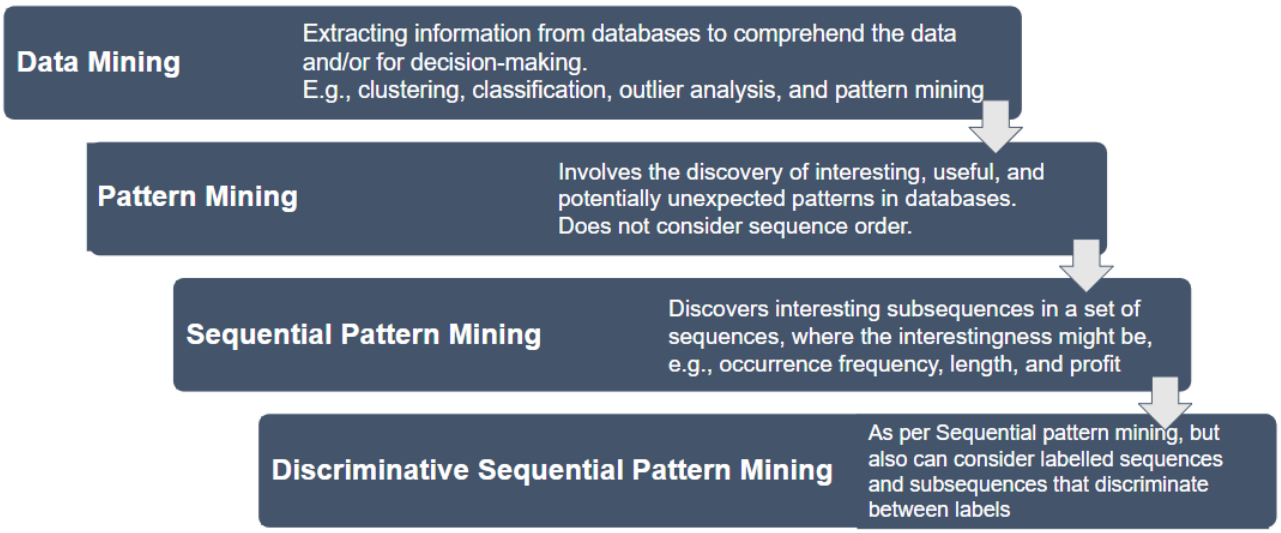


Figure 6: Discriminative sequential pattern mining within the field of data mining. *This figure is adapted from Fournier-Viger et al. (2017) to include the sub-discipline of discriminative sequential pattern mining.*

2.3 Sports Performance Analytics

Sports analytics and sports performance analysis developed historically in parallel but largely independently. As a result, there are differences in the terminology used between sports analytics and sports performance analysis. For example, in sports performance analysis, performance indicators (Hughes & Bartlett, 2002; Sampaio & Leite, 2013) might be referred to as variables or covariates, whereas they might be referred to as predictors or features in sports analytics. The techniques commonly used in each discipline have also traditionally differed. As mentioned, sports performance analysis has historically emphasised statistical methods, whereas sports analytics branches out more into methods from computer science and data science. Finally, the journals in which studies are published

(and other dissemination venues) have traditionally differed. For instance, the International Journal of Performance Analysis of Sports is a well-known journal related to sports performance analysis, while journals that are more focused on sports analytics include the Journal of Sports Analytics and the Journal of Quantitative Analysis in Sports.

In recent years, the disciplines have become more intertwined through the incorporation of advanced analytics techniques traditionally employed primarily in sports analytics into sports performance analysis. There could be benefits in sports performance analysis evolving along with the performance-related side of sports analytics (i.e., not the sports management analytics side) into a sports performance analytics discipline.

2.3.1 Studies in sports performance analytics

This sub-section outlines some recent studies that appear to lie at the intersection of sports performance analysis and sports analytics, which, it is argued, is emerging as a distinct discipline, “sports performance analytics”, which combines sports performance analysis with sports analytics. There is a growing connection between these two fields. The studies presented in this thesis also aim make contributions to the emerging discipline of sports performance analytics.

An early sports performance analysis study published in sports performance analysis speciality journal, the International Journal of Performance Analysis of Sports — which considered a problem more commonly considered in sports analytics, match result prediction using machine learning — was that of Reed & O'Donoghue (2005). The authors applied seven different predictive models, including multiple linear regression, ANNs, and expert predictions, to seven variables derived from English Premier League soccer and Premiership Rugby Union matches. Surprisingly, soccer was found by the authors to be able to be predicted with higher accuracy than rugby, and it was found that the predictive models were able to outperform the human experts. The authors mentioned that it was the first time that AI models could be said to outperform experts. Notably, this study referred to "predictive models of performance", thus linking predictive models, which are more commonly used in sports analytics, and sports performance analysis.

As mentioned previously, contextual (situational) variables such as match location, officials, stage of the season, and opposition quality can be significant since, although these variables

are not under the control of teams or players, they can affect performance. They can thus be helpful to control for. Parim, Güneş, Büyüklü, and Yıldız (2021) used contextual variables and performance indicators to predict group stage Champions League soccer, employing a one-way Analysis of Variance (ANOVA) and the Tukey honestly significant difference test to determine key performance indicators that were most relevant to match results. K-means clustering was used to group opposition teams into one of three quality groups: weak, balanced, and strong, and multidimensional scaling and decision trees were then applied to each of these three clusters. The authors found that while some of the performance indicators of teams varied based on opposition quality, specific performance indicators increased win probability regardless of the quality of the opposition team. Bilek and Ulas (2019) investigated the effect of contextual variables and performance indicators on match outcomes in the 2017-18 English Premier League season, based on opponent quality, using statistical and machine learning methods. While the most influential variable affecting match outcome in the decision tree model was scoring a goal first, incorporating opposition quality was necessary since the effects of many predictors on match outcome varied according to the quality of the opposition team.

As highlighted in some of the publications that comprise this thesis, interpretability is vital to decision-makers in sporting contexts. Robertson, Back, & Bartlett (2016) considered two seasons of Australian Football League (AFL) Australian Rules Football, analysing the extent to which typical team performance indicators in relative form (i.e., standardised against the opposition team) could explain match outcomes. Logistic regression and chi-squared automatic interaction detection (CHAID) decision trees yielded similar performance. However, the CHAID decision trees model identified multiple winning performance indicator profiles, making it more useful in a practical context for strategic planning for upcoming matches. CHAID's interpretability was noted to be an appealing aspect for application in sporting contexts. Parmar et al. (2017) utilised a variation of CHAID decision trees called Exhaustive CHAID decision trees. The authors analysed 24 performance indicators, representing the difference in home and away team values, from three seasons of European Super League Rugby League data obtained from Opta. Backwards logistic and linear regression models for discrete match outcome and numeric points difference prediction were employed in conjunction with Exhaustive CHAID trees to identify key performance indicators. Another interpretable decision tree model incorporating boosting is the ADTree (Freund & Mason, 1999), employed in Bunker et al. (2023). Furthermore, Bunker & Spencer (2022) and Bunker et al. (2021) generated interpretable and practically-applicable results.

Another recent area of enquiry is investigating the impact of player-level performance variables on team performance. Measures such as plus-minus ratings analyse what happens in terms of goals scored while a particular player is on the field and when they are not (Kharrat, McHale, & Peña, 2020; Hvattum, 2020). Apostolou and Tjortjis (2019) predicted player positions and used historical data to estimate the goal-scoring performance of a player in the following season. The authors also predicted the number of shots a player makes in each match. Li, Ma, Gonçalves, Gong, Cui, & Shen (2020) built a Linear Support Vector Classifier (LSVC) model for predicting team ranking and match outcomes in Chinese Super League soccer, engineering 22 features related to performance in terms of offence, passing, and defence. The LSVC model was used to compute a team performance ranking, and feature weights were used to determine the features most relevant to match the outcome. Pantzalis & Tjortjis (2020) considered the performance of teams and players in soccer, aiming to predict the final league team ranking table for specific leagues and the performance of teams relative to the most recent prior season. Yücebaş (2022) noted a general lack of studies that have used traditional statistical or machine learning techniques to investigate the impact of players' performances on match outcomes and proposed using deep learning for this purpose.

Deep learning models such as convolutional neural networks (Hsu, 2021) and long short-term memory (LSTM) networks (Zhang et al., 2022) have been increasingly employed in sports in recent years. Deep learning has also been used for the analysis of sequential data in soccer to predict subsequent events that will occur in matches (Simpson, Beal, Locke, & Norman, 2022; Yeung, Sit, & Fujii, 2023) as well as the prediction of match results in soccer (Yeung, Bunker, Umemoto, & Fujii, 2024). Wang et al. (2020) identified the need for concurrent match outcome prediction and performance evaluation. They proposed generating a forward-looking "win rate curve" instead of predicting discrete wins and losses, with actions resulting in changes in the position of this curve. A deep learning model incorporating a recurrent attention mechanism and matrix perturbation was leveraged to learn and generate this win rate curve, with the model encoding player behaviour and capturing interactions between players, meaning performance could be quantified at different levels (e.g., player and team levels).

2.3.2 A proposed sports performance analytics framework & workflow

Overall, there is evidence of an area of enquiry developing that represents an evolution of sports performance analysis into sports performance analytics[6] in which advanced analytics techniques traditionally used in sports analytics are being utilised in conjunction with domain knowledge from sports performance analysis being incorporated, for example, through the use of performance indicators within predictive models.

As previously mentioned, Link (2018) noted that sports analytics can be partitioned into sports management analytics and sports performance analytics. Thus, one view may be that sports performance analytics simply organically grew out of sports analytics in isolation. However, it is more likely that conceptual aspects of sports performance analysis are being incorporated into sports analytics, as demonstrated by the publications covered in the previous subsection. Another view is, therefore, that sports performance analytics is a discipline evolving out of both sports performance analysis and sports analytics, incorporating aspects of both disciplines (Figure 7).

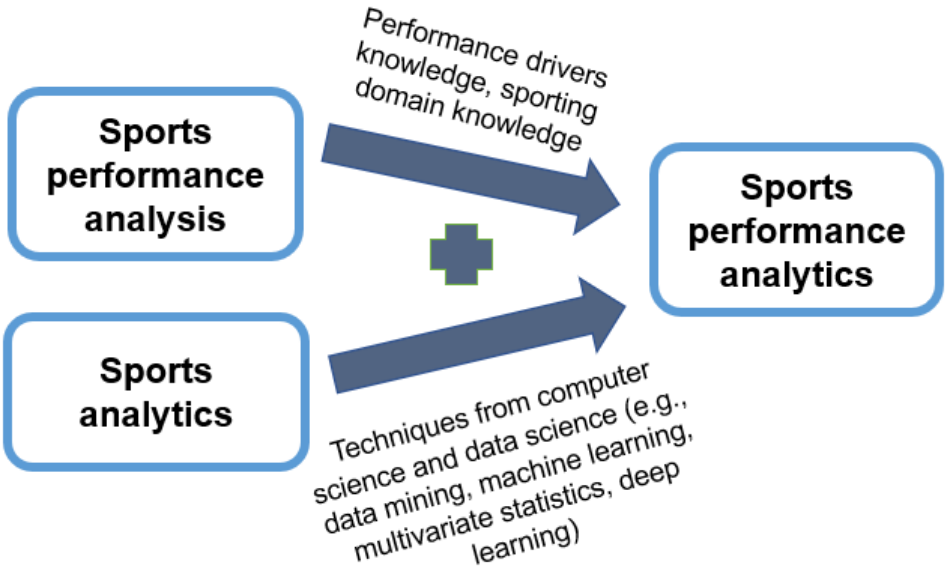


Figure 7: Sports Performance Analytics. *Sports performance analytics is an evolution of sports performance analysis and sports analytics, combining concepts from the former with methods from the latter.*

Other evidence for the emergence of an emerging discipline is that some professional teams and university programmes have used the term “sports performance analytics”. For

[6] Another interpretation is that sports performance analytics could be a combination of sports analytics and sports performance analysis (i.e., at the intersection of these two disciplines).

instance, Lorena Martin, who authored a textbook in the domain (Martin, 2016), developed a course in Sports Performance Analytics at Northwestern University in the United States and formerly held a position as Director of Sports Performance Analytics at the Los Angeles Lakers (Martin, 2015; Martin, 2019). Furthermore, in 2016 and 2018, two books were published that contained “sports”, “analytics”, and “performance” in their titles (Martin, 2016; Jayal et al., 2018).

Miller (2015) mentions “sports performance analytics” specifically but also suggests it is relevant in sports management, for example, in terms of having appropriate information systems in place and in taking relevant measurements so that the numerous factors that contribute to successful on-field or on-court outcomes can be comprehended. In Link’s (2018) definition of sports analytics as “the process of searching, interpreting and processing information in sports-related performance data using information systems and mathematical methods of data evaluation to achieve competitive advantages”, the presence of the word “performance” is notable. In contrast to Miller (2015), Link notes the distinction between sports performance analytics and sports management analytics.

A potential workflow/process for sports performance analytics is now proposed, combining aspects of sports analytics and sports performance analysis (Figure 8). Incorporating the earlier discussions of sports analytics and sports performance analysis, sports performance analytics could be described as the process of augmenting and analysing event data, spatiotemporal tracking data, and other performance-related data, taking into account contextual factors and using advanced analytics techniques to analyse past performance or predict future performance in competition or training. Sports performance analytics also involves utilising appropriate information systems and visualisations to inform decision-makers. As with sports performance analysis and sports analytics, the goal of sports performance analytics is to improve performance by enhancing competitiveness and gaining a strategic on-field advantage. Whether performance has improved can then be assessed, and if not, the data collection process can be refined (e.g., additional or different data may need to be collected). Other performance-improving actions could be using different analytics techniques or adjusting output dissemination to decision-makers such as coaches.

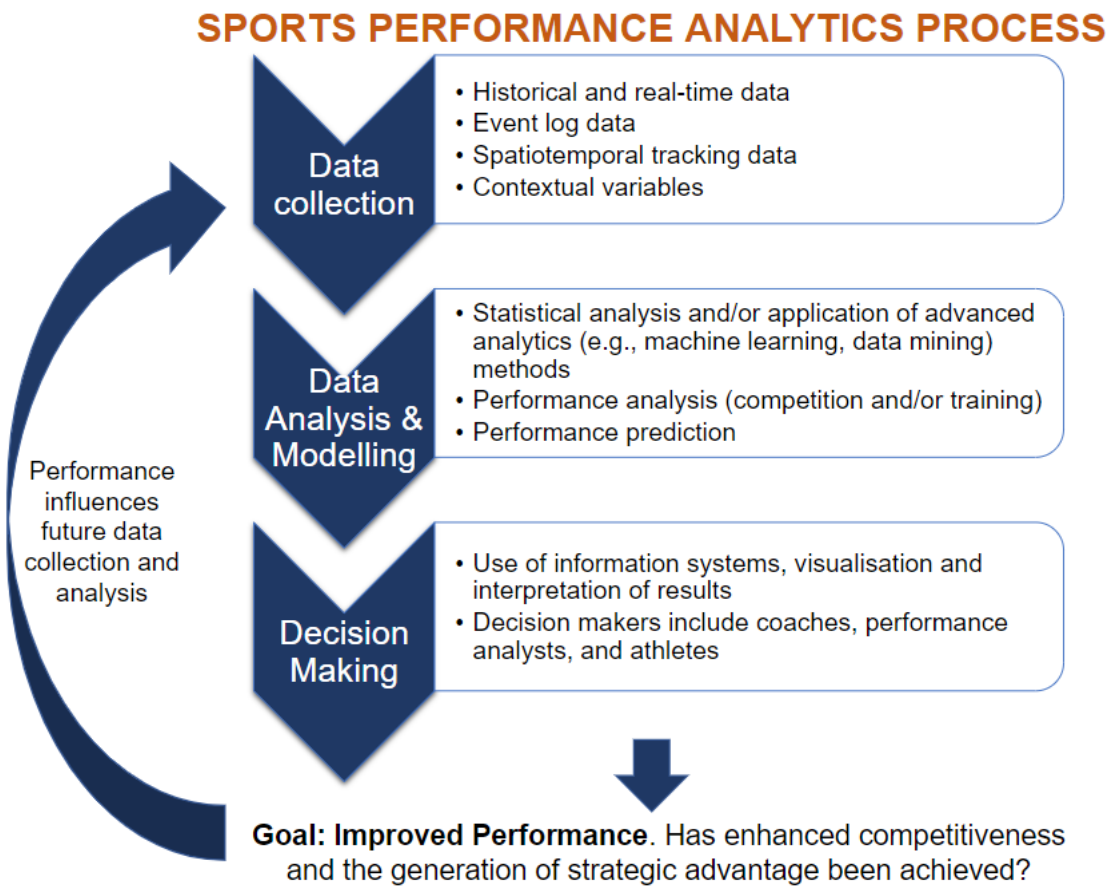


Figure 8: Sports Performance Analytics Workflow. *A proposed sports performance analytics workflow that incorporates aspects of both sports analytics and sports performance analysis.*

As a means of narrowing the scope of the field, it could be assumed that the actions performed by individual players, which lead to on-field events, is the lowest level of analysis considered in sports performance analytics. At a lower level, physiological factors and technique-related movements are also important in leading to effective actions — and machine learning can indeed be used to analyse these (Sweeting, Ball, & Robertson, 2018). However, the analysis of physiological factors and technique often requires significant sporting domain expertise, which would limit the accessibility of the sports performance analytics field to only researchers who possess significant sports science or biomechanical knowledge in addition to expertise in computer science and data science related techniques. Thus, it is proposed that these lower levels could fall outside the scope of sports performance analytics and instead be considered in sports science, sports informatics, or even sports engineering.

2.3.3 A multi-level approach to sports performance analytics

This thesis takes a multi-level approach to sports performance analytics in the sense that the studies investigate performance at different levels of analysis, for example, at the match outcome, tournament stage, and passage of play levels. In the two studies presented in the appendix, performance is investigated at the passage of play level (using spatiotemporal data rather than event data) and at the player pair level, representing all historical match results between two players in tennis.

Figure 9 depicts a hierarchy showing some levels at which performance can be analysed in sports. The figure also shows the studies in this thesis that considered performance at that level. Starting from the lowest level in Figure 9, individual players perform actions that result in on-field events tagged in notational systems, which generate event data. As mentioned, this level of analysis is the most granular considered in sports performance analytics. In addition, on- and off-the-ball player movements can be analysed using spatiotemporal tracking data, which can be augmented with event data for more holistic analyses.

Matches, in turn, make up league or competition rounds (or tournaments/tournament stages), which then make up seasons. A gauge of success at a season level can be determined by observing the ratio of wins to losses a team achieved over a particular season. For example, Bayer Leverkusen in the 2023/24 Bundesliga, the Golden State Warriors in the 2015/16 season, and Arsenal in the 2003/04 English Premier League season had exceptional win-to-loss ratios (GQ, 2024). Performance over several seasons or years is indicative of the long-term performance of a team or athlete. Remarkably, teams or individuals can dominate a sporting competition over many years — even at the professional level. One team that comes to mind is Bayern Munich, who won 11 Bundesliga competitions before the 2023-24 season.

In team sports, there is another parallel hierarchy related to personnel (see the grey tiles on the left side of Figure 9). The actions and movements of players and subsequent on-field events determine performance at the player level, and passages of play make up periods of play that involve groups of players. In team sports, players form part of units or groups within teams. For instance, there may be midfield, defensive, and offensive groups in soccer. In American Football, offensive and defensive teams switch between plays; therefore, offensive and defensive units would be of interest as player groups to analyse performance. In Rugby Union, a more appropriate division of players into groups could be into forwards

and backs, or more granularly, loose-forwards, locks, and props (forwards), halves (halfback and first-five-eight), midfielders (second-five-eight and centre), and outfielders (wingers/fullback).

Then, match outcomes make up rounds of competitions/stages of tournaments and then seasons. Season-level performance largely depends on team-level performance since squads are fixed during the season. Teams that achieve sustained strong performance over multiple seasons must first ensure that the foundation was laid in performance at the more granular player and player group levels. Furthermore, team rosters change over multiple seasons; thus, club culture becomes essential in achieving continued success over several seasons. However, success in this context is perhaps venturing into concepts explored in sports and organisational psychology (Cole & Martin, 2018).

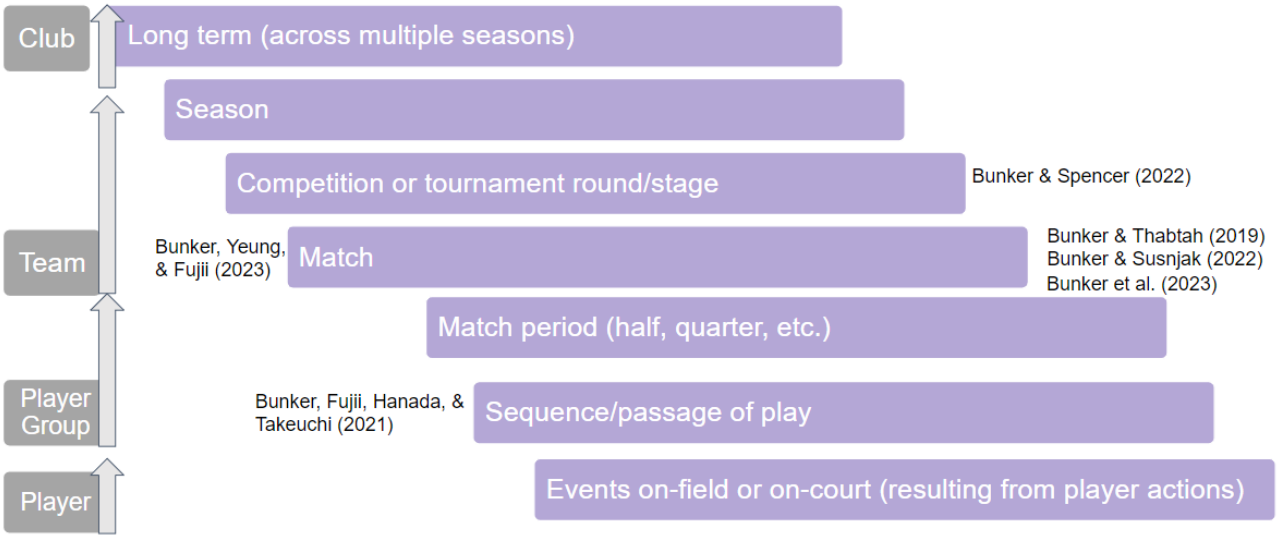


Figure 9: Levels at which performance can be analysed. *A hierarchy depicting some levels at which performance can be analysed in team sports, and the corresponding studies in this thesis relate to these performance levels. Also shown in grey is a separate, additional hierarchy that relates to the personnel who influence each level of performance. Player actions leading to events are assumed to be the lowest performance level analysed in sports performance analytics.*

CHAPTER 3: CONNECTIONS BETWEEN PUBLICATIONS & CONTRIBUTIONS OF THE BODY OF WORK

This chapter describes the ways in which the publications comprising this thesis are connected and discusses how the body of work constitutes a cluster of original research. In doing so, the chapter also emphasises the key contributions of the thesis overall, and how it addresses the research questions specified in subsection 1.1. Each study also has its own set of contributions; please see the publication enclosed in the corresponding chapter (Chapters 4 to 9). The studies in this thesis are conceptually connected in two primary ways, which are depicted in Figure 10. First, all six studies investigate machine learning in some capacity. Second, in some cases at different levels of analysis, all studies investigate performance in sports. In terms of how one study led to another, the publications in this thesis are also connected to each other chronologically (see the top panel in Figure 11).

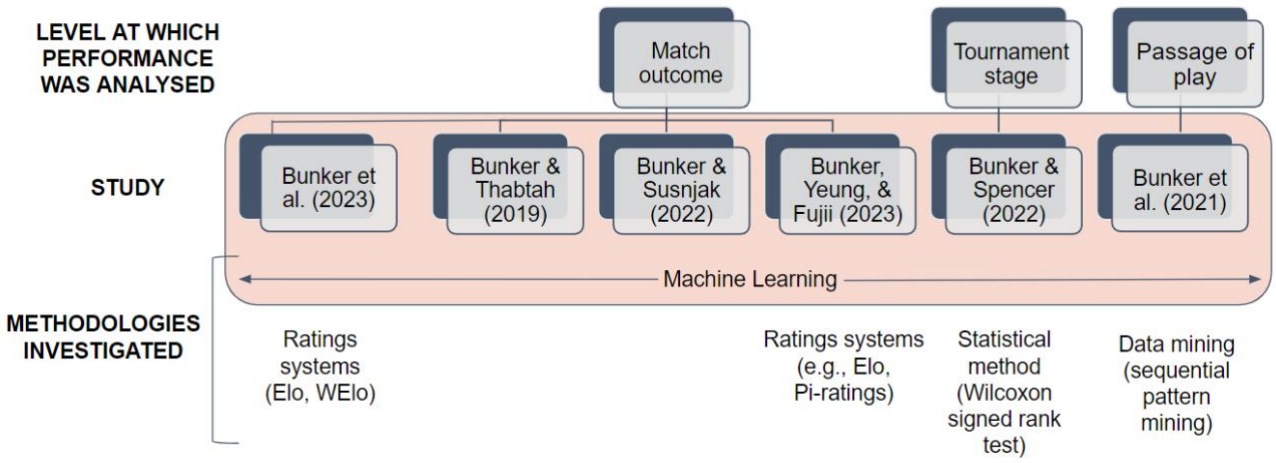


Figure 10: Methodological Publication Connections. *The six publications in this thesis all investigate performance, with the level at which performance was analysed differing. The methodologies employed differed among studies, but the commonality among the six studies was that all publications investigated machine learning.*

Table 1 presents the studies that make up this thesis, their DOIs, the type of study, the level at which performance was analysed, and the methodologies employed in the study. Four studies in the thesis investigated sports match outcome prediction using machine learning (Bunker & Thabtah, 2019; Bunker & Susnjak, 2022; Bunker, Yeung, Susnjak, Espie, & Fujii, 2023; Bunker, Yeung, & Fujii, 2025); that is, they considered performance at the match level.

The remaining studies considered performance at the passage of play (Bunker et al., 2021) and tournament stage levels (Bunker & Spencer, 2022).

***Table 1**: **Publications in the main text of the thesis.** Studies contained in this thesis, along with their DOIs, the level at which performance was analysed, and the methodologies employed and/or compared in the study.*

Study Title	**DOI/URL**	**Study Type**	**Level at which performance was analysed**	**Methodologies considered**
Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied computing and informatics*, 15(1), 27-33.	https://doi.org/10.1016/j.aci.2017.09.005	Conceptual Framework, Synthesis	Match outcome	Machine learning
Bunker, R., Fujii, K., Hanada, H., & Takeuchi, I. (2021). Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: an application to rugby union. *PLOS One*, 16(9), e0256329.	https://doi.org/10.1371/journal.pone.0256329	Application of an existing method in the context of sports	Passage of play	Discriminative sequential pattern mining/machine learning
Bunker, R. P., & Spencer, K. (2022). Performance indicators contributing to success at the group and play-off stages of the 2019 Rugby World Cup. *Journal of Human Sport and Exercise*, 17(3), 683-698.	https://doi.org/10.14198/jhse.2022.173.18	Application of an existing method in a new sporting context	Tournament stage (group & playoff)	Rules-based Machine learning, statistical techniques
Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, 73, 1285-1322.	https://doi.org/10.1613/jair.1.13509	Review and Synthesis	Match outcome	Machine learning
Bunker, R., Yeung, C.., Susnjak, T., Espie, C., & Fujii, K. (2023). A comparative evaluation of Elo ratings- and machine learning-based methods for tennis match result prediction. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*	https://doi.org/10.1177/17543371231212235	Comparative evaluation, application of an existing method in the context of sports	Match outcome	Machine learning, Ratings
Bunker, R., Yeung, C., Fujii, K. (2025). Machine Learning for Soccer Match Result Prediction. In: Blondin, M.J., Fister Jr., I., Pardalos, P.M. (eds) *Artificial Intelligence, Optimization, and Data Sciences in Sports*. Springer Optimization and Its Applications, vol 218. Springer, Cham. https://doi.org/10.1007/978-3-031-76047-1_2	https://doi.org/10.1007/978-3-031-76047-1_2	Review and synthesis	Match outcome	Machine learning

Although widely applied in many other machine learning and data mining application domains, the CRISP-DM framework proposed by Shearer (2000) had not been considered to a wide degree in the context of sports match outcome prediction. One exception was the study of Delen, Cogdell, & Kasap (2012). With the thought of the need for a structured methodological and experimental approach specific to sports match outcome prediction provided by such frameworks, Bunker & Thabtah (2019) developed a framework for machine learning for sports match result prediction. The Sports Result Prediction CRISP-DM (SRP-CRISP-DM) framework, which extended the CRISP-DM framework, was proposed after conducting critical analysis and synthesis of the literature related to the application of ANNs in this domain (RQ1). ANNs were commonly used models in early studies in the domain, often as the sole model. In the literature survey prior to the presentation of the SRP-CRISP-DM framework itself, both individual sports as well as team sports were considered. The proposed Sports Result Prediction CRISP-DM framework (SRP-CRISP-DM) was devised by adapting the CRISP-DM framework (Shearer, 2000) for the specific challenges faced when predicting match results using machine learning. Some of these challenges include the temporal ordering of matches (which makes traditional cross-validation, which shuffles instances randomly, inappropriate), partitioning temporally-ordered match-, season-, and round-level data into training, validation, and test sets, and the selection and engineering of relevant feature subsets as input to machine learning models (e.g., match features derived from events that occur within matches, betting odds, expert-selected features, and features that are external to a match and are not derived from match events such as match venue, and match officials).

Leading on from Bunker & Thabtah (2019), it was decided that a critical analysis and synthesis of the literature on machine learning for sports match result prediction, which was not just restricted to ANNs, would be a valuable contribution to the literature. In particular, Bunker & Susnjak (2022) had a broader focus than Bunker & Thabtah (2019) in terms of the types of machine learning models used in the studies covered (all types of machine learning models were considered rather than just ANNs), but a more narrow scope in terms of the types of sports (team sports were covered). Indeed, a recommendation uncovered in Bunker & Susnjak (2022) was that future researchers consider a set of candidate models rather than only relying on ANNs. Bunker & Susnjak (2022) covered the characteristics of different sports that can affect their predictability and recommended that future researchers consider these characteristics when evaluating the performance of models. Bunker & Susnjak (2022) surveyed selected studies published in machine learning for match outcome prediction

domain between 1996 and 2019. The review considered team sports, including invasion and striking/fielding sports (see Figure 12 in Appendix A1 for categorising sport types). The survey identified machine learning algorithms commonly applied in the domain, datasets, accuracies achieved and approaches for evaluating models. It also discussed future research avenues and recommendations for further work in the domain. The study identified the potential for a higher degree of interdisciplinary collaboration between researchers from sports science (and sports performance analysis, more specifically) and those from machine learning, especially for selecting and engineering a set of relevant features for the machine learning model. Nearly 90% of the studies surveyed in Bunker & Susnjak (2022) mentioned engineering additional, more informative features as a direction for future research. The study also identified feature engineering as being of greater importance for model performance in the domain than having a large dataset in terms of number of matches. This finding is consistent with what Domingos (2012) highlighted for machine learning in general: using domain knowledge to generate more informative and discriminative features is generally the best approach to improving predictive performance. Finally, we discussed the inherent difficulty in predicting certain sports due to their characteristics, for example, whether a sport is low-scoring or high-scoring, how the score is incremented, the competitive balance of the sport, and the propensity for draws.

Subsequently, based on its global popularity as well as the identification in Bunker & Susnjak (2022) of inherent characteristics of soccer that make it particularly challenging to predict, the book chapter Bunker, Yeung, & Fujii (2025) had a more specific focus in terms of only considering soccer. This narrower scope allowed for a more in-depth treatment. Conducting survey papers in the domain of machine learning for sports match outcome prediction is becoming increasingly challenging because of the explosion in the number of papers being published, especially over the past decade or so. This became apparent while writing Bunker & Susnjak (2022). It is, therefore, suggested that researchers considering writing survey papers in the domain focus on a specific sport.

As well as the development of novel methods, there is significant scope for the utilisation of machine learning and data mining techniques, which have been developed for use in other domains, in sporting contexts. This can be achieved, for example, by examining the structure of data to which such methods are being applied in other domains, and investigating which data produced in sports exhibits a similar structure. In many domains including sports, methods that are interpretable are often useful, or even vital, for deciphering the underlying

logic behind a model's predictions. Leveraging interpretable methods from other domains or other contexts, for example, for machine learning for match outcome prediction (Bunker et al., 2023) and key pattern identification (Bunker et al., 2021; Bunker & Spencer, 2022; Bunker et al., 2024 - appendix), and demonstrating their utility in a sporting context and with a focus on interpretability such that results and/or models could be helpful in practice, was a core focus of this thesis (RQ2/RQ3).

The sequential nature of event data captured in video analysis systems makes data mining and, more specifically, sequential pattern mining approaches, appropriate analytical tools. Sequential pattern mining techniques have been used in fields such as market basket analysis in retail, for example, to identify customer purchase patterns (e.g., Goel & Mallick, 2015), and for analysing biological sequences (Wang, Xu, & Yu, 2004). Discriminative sequential pattern mining methods can be suitable when such sequences can be assigned labels, and when the goal is to identify subsequences that discriminate between the labels. A supervised discriminative sequential pattern mining technique called the S3P-classifier (Nakagawa et al., 2016; Sakuma et al., 2019) — had been applied to animal trajectories converted to sequences (Sakuma et al., 2019) but not previously been employed in sporting contexts — was leveraged in Bunker et al. (2021) (RQ2). Performance was considered at the passage of play level, and the S3P-classifier was applied to labelled event sequences derived from event log data. The data were delimited, with the beginning and end of each passage of play/event sequence determined based on specified criteria (see the paper itself for further details). Labels were then assigned to each sequence based on the terminal outcome of the passage of play according to whether the sequence of play led to points being scored[7]. From the labelled passage of play sequences, the method identified key patterns as subsequences of events that discriminated between scoring and non-scoring outcomes. Compared to unsupervised sequential pattern mining methods that had previously been applied in sports (Hrovat et al., 2015; La Puma & de Castro Giorno, 2017; Decroos, Van Haaren, & Davis, 2018), the study found that the supervised discriminative sequential pattern mining method identified patterns with a greater diversity of events that would be of greater value to coaches when interpreted. The passage of play level is, of course, a more granular level of performance than match outcome since matches in invasion sports generally consist of a series of passages of play.

[7] In the case of penalty goals, a shot being attempted was deemed to be sufficient since a scoring opportunity was still created and whether it was successful is based on kicking skill and form.

Major tournaments in team sports such as soccer and rugby often consist of group (or pool) stages, followed by playoff matches. It seems plausible that teams may need to adjust their playing strategies from one stage of the tournament to the next. For instance, effective defensive play may become more important at the playoff stage of a tournament. An important concept in sports economics is competitive balance (Zimbalist, 2002), and various measures of competitive balance can be computed to determine the degree to which a competition is even or uneven (Humphreys, 2002). International rugby is not as competitively balanced as international soccer, so some one-sided contests generally occur at the group stage of the Rugby World Cup tournament. Bunker & Spencer (2022) investigated performance at the tournament stage level, specifically the group and playoff stages of the 2019 Rugby World Cup. The tournament stage level is a more aggregated level of performance than the match level of analysis since outcomes at the tournament stage level reflect individual match results at a specific tournament stage. Progression to the playoff stage of a tournament is commonly based on points differential, which is some function of points gained as a result of wins and points scored in group-stage matches. The rules-based machine learning algorithm RIPPER (Cohen, 1995) had previously been used in fields such as text classification (Cohen & Singer, 1996), malware detection (Dolejš & Jureček, 2022), phishing detection (Thabtah & Kamalov, 2017), and for feature selection for match outcome prediction in basketball (Thabtah, Zhang, & Abdelhamid, 2019). Some of the advantages of RIPPER include its speed, but more importantly, it generates interpretable decision rules. In Bunker & Spencer (2022), RIPPER was used alongside the traditionally utilised statistical technique in this type of study (Wilcoxon signed-rank test), which was applied to each feature univariately. RIPPER could identify interpretable key patterns in the form of decision rules composed of performance indicators and their values that discriminated between winning and losing at the group and playoff stages of the tournament (RQ3). Unlike the Wilcoxon signed-rank test, RIPPER was also able to identify combinations of performance indicators that are associated with success, not just identifying discriminative performance indicator variables in a univariate manner. There appears to be scope in future research to further investigate the use of multivariate methods, not only statistical but also machine learning based multivariate methods, which can identify pertinent combinations of performance indicators contributing to success rather than analysing the importance of each performance indicator in isolation as univariate methods do. Provided the models or results of the models are interpretable, these types of methods could have considerable value in practice for coaches and performance analysts.

One way in which machine learning and data mining techniques will find greater application for assessing performance in sports is to ensure that models and/or results are interpretable, so that coaches and performance analysts are able to use the insights gleaned in practice. In Bunker et al. (2023), a comparative evaluation of Elo ratings and Weighted Elo (WElo) (Angelini, Cantila, & De Angelis, 2022) ratings with machine learning models was conducted for match outcome prediction in tennis. This study is connected to some of the other studies in this thesis not only because it proposes an interpretable approach to machine learning for sports match result prediction (RQ3) but also because the SRP-CRISP-DM framework that had been proposed in Bunker & Thabtah (2019) was demonstrated in this study in practice to guide the study. Alternating Decision Trees (ADTree) (Freund & Mason, 1999) is a boosted tree model that had previously been used in domains where interpretability is essential, for example, in predicting disease in medicine (Liu, Lin, Zhou, & Wong, 2005; Jabbar, Deekshatulu, & Chndra, 2014) and in landslide susceptibility modelling (Pham, Tien Bui, & Prakash, 2017; Wu et al., 2020). Despite incorporating the accuracy-enhancing benefits of boosting (the AdaBoost boosting algorithm (Freund & Schapire, 1997) is used to grow the Alternating Decision Tree), the ADTree model retains an interpretable tree structure (similar to CHAID, which was described earlier in this contextual statement). ADTrees had yet to be previously applied in sporting contexts despite interpretability being important to decision-makers such as coaches. In Bunker et al. (2023), ADTrees were used as one of a set of machine learning models whose performance was evaluated against Elo and Weighted Elo ratings for tennis match result prediction. ADTrees were shown to achieve robust performance, and their interpretability for determining the effects of particular features on match outcome was an appealing feature compared to machine learning models that are black box[8] in nature. ADTrees could find application in other sporting contexts due to their interpretable model structure. Furthermore, these types of comparisons of machine learning with ratings based methods are interesting as ratings can be used as predictive methods themselves (taking the team/player with the higher computed rating as the winner) but also as features within models. Thus, it is interesting to consider ratings in each of these capacities, and to compare their performance with machine learning methods and whether ratings can improve machine learning methods' performance when used as model features.

Leading on from Bunker & Susnjak (2022), the book chapter Bunker, Yeung, & Fujii (2025) had a more specific focus: the use of machine learning for match outcome prediction in

[8] Black box models such as deep neural networks produce predictions without giving clear insight into their internal decision-making processes (Karim et al., 2023)

soccer. As mentioned, we discussed in detail in Bunker & Susnjak (2022) the inherent difficulty of predicting match results in certain sports. Soccer is an exciting sport to attempt to predict not only because it is the most widely played and followed sport globally but also because it is a sport that has specific characteristics that make it inherently difficult to predict. For instance, it is low-scoring, often has highly competitive leagues (competitively balanced), and a draw is a common outcome. The chapter discussed datasets, models, features, and model evaluation methods in machine learning for match outcome prediction in soccer in detail, providing synthesised insights and findings (RQ1). The chapter also aimed to provide future researchers with a broad overview of existing research and potential avenues for further work. The work identified the current state-of-the-art approaches in the domain — at least on datasets containing goals as the only features related to on-field events — as those that use soccer-specific rating systems such as pi-ratings (Constantinou & Fenton, 2013) and Berrar ratings (Berrar, Lopes, & Dubitzky, 2019) as model features in gradient boosted tree models (Razali, Mustapha, Mostafa, & Gunasekaran, 2022; Hubáček, Šourek, & Železnỳ, 2019b) such as XGBoost (Chen & Guestrin, 2016) or CatBoost (Prokhorenkova et al., 2018). Another key observation from the chapter was that a comparative evaluation of the current state-of-the-art gradient-boosted tree models against more recently proposed deep learning models, for example, deep neural networks (Rahman, 2020) and those specifically designed for time series type data such as Transformers (Vaswani, 2017) (Simpson, Beal, Locke, & Norman, 2022; Yeung, Sit, & Fujii, 2023) and attention-based LSTM (Zhang et al., 2022), should be explored in further research. This will help clarify whether, ideally, on a range of datasets, the state-of-the-art approach for machine learning for soccer match outcome prediction remains gradient-boosted trees or whether deep learning could supplant these models. Finally, the chapter discussed the need for the interpretability of models to be enhanced in order for them to become useful for coaches and sports performance analysts rather than merely for prediction purposes.

In summary, the publications in this thesis that provided conceptual frameworks and critical analysis and synthesis in the domain of machine learning for sports match outcome prediction (top panel, Figure 11) were primarily linked in a chronological sense, and the scope and focus of these studies changed over time as the domain was investigated in more detail. Both Bunker, Yeung, and Fujii (2025) and Bunker and Susnjak (2022) carried out critical analyses of existing literature, deriving relevant insights in the field as well as opportunities for further research (RQ1). While Bunker & Thabtah (2019) found that ANNs were heavily used in early studies investigating match outcome prediction using machine

learning, often as the sole model, Bunker & Susnjak (2022) noted that the utilisation and comparative evaluation of a set of candidate machine learning models is a preferable approach. The scope of the studies changed over time. Bunker & Susnjak (2022), by focusing on team sports in general, including both invasion and striking/fielding sports, had a broader coverage and discussed in detail the inherent characteristics in different sports that can make match outcome prediction more challenging in certain sports. Following this, Bunker, Yeung, and Fujii (2025) focused on a specific invasion sport, soccer, which has inherent characteristics that make it challenging to predict.

This work aims to bridge the gap between match outcome prediction in sports and sports performance analysis. As mentioned, in the review paper on machine learning for team sports match outcome prediction (Bunker & Susnjak, 2022), we highlighted the potential benefits of greater inter-disciplinary collaboration between machine learning researchers and researchers/practitioners in the sports science domain, particularly concerning the selection and engineering of relevant model features to be used as input for sport results prediction machine learning models. This is effectively the incorporation of domain knowledge into the modelling process, which has been known to boost performance in sports prediction using machine learning (Joseph, Fenton, & Neil, 2006; Berrar, Lopes, & Dubitzky, 2019; Berrar, Lopes, & Dubitzky, 2024). Incorporating domain knowledge into the modelling process through, for example, feature selection is one thing; however, to bridge the gap between sports outcome prediction using machine learning and the analysis of performance in sports, models and/or results should be interpretable.

This thesis has explored a few approaches for incorporating interpretability (RQ3). First, the results from data mining and machine learning models can be interpretable. For instance, the discriminative event subsequence results from Bunker et al. (2021) were readily interpretable, as were the decision rules of performance indicators RIPPER obtained by Bunker & Spencer (2022). Another approach for interpretability is generating a model that is interpretable in and of itself (e.g., the ADTree model). Another way interpretability can be incorporated, discussed in Bunker, Yeung, & Fujii (2025), is to employ explainable AI techniques to decipher the effects of particular model features on target variables from black box models. This will be an increasingly important area of enquiry as deep learning models become increasingly employed in sports performance analytics, and for the comparison with traditional machine learning methods not only in performance but also in terms of interpretability of model outputs.

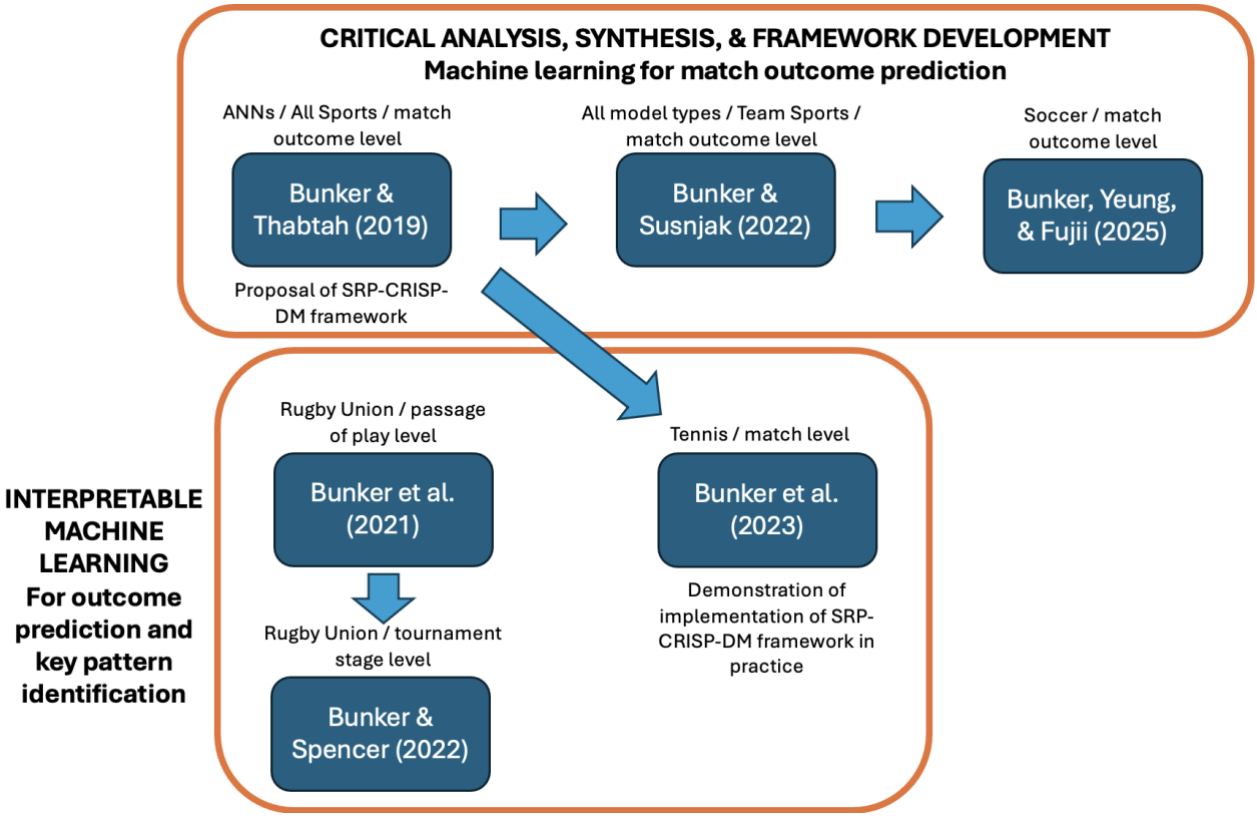


Figure 11: Interconnectedness of the publications contained in this thesis. *This figure shows the methods investigated, sport(s), and level of performance considered. The studies in the top panel are connected in a chronological sense and also in the narrowing of focus and scope. The framework proposed in Bunker & Thabtah (2019) (top panel) was demonstrated in practice in Bunker et al. (2023) (bottom panel). Studies in the bottom panel are related in terms of their focus on leveraging methods from other domains and contexts in sports, and two studies investigated rugby union using different types of data.*

CHAPTER 4: PUBLICATION 1 “A MACHINE LEARNING FRAMEWORK FOR SPORT RESULT PREDICTION”

This study investigated performance at the match outcome level, conducting critical analysis and synthesis of literature related to machine learning for sports result prediction, focusing on the application of Artificial Neural Networks (ANNs), which were a commonly applied machine learning technique in the domain in early studies in particular (often as the sole model). The paper described the datasets utilised, model evaluation methods, and limitations of existing studies. The study then proposed a practical, conceptual framework called Sport Result CRISP-DM (SRP-CRISP-DM) to guide the use of machine learning techniques for sports match result prediction. SRP-CRISP-DM builds on the CRISP-DM framework (Shearer, 2000) and accounts for the specific factors that must be considered when utilising machine learning models to predict match outcomes in sports.

The main components of the framework include, first, understanding the problem and the objective of a model (e.g., to outperform experts, to compete in a competition, to place bets to make a profit, or to understand factors contributing to winning), along with the characteristics of the sport and factors that are likely to influence the outcome, for example, through a literature survey, or by leveraging expert or personal domain knowledge. It is then necessary to understand the available data, how the target variable to be predicted will be defined (e.g., numeric or discrete), and the granularity of the data, for example, whether both player- and team-level data is available or only team-level data. The next step involves preprocessing the data, as well as engineering and selecting relevant features, creating different subsets of features to compare and computing historical averages of features where applicable (i.e. when a particular feature is not known until after a match has been played, which is the case for variables derived from match events). Candidate models should then be selected based on a literature survey and compared by evaluating their performance using an appropriate performance measure on different temporally ordered data partitions into training, test, and validation sets. The best-performing model can be deployed for prediction in a production environment if necessary.

In the context of tennis match result prediction, Bunker et al. (2023) (Chapter 8) subsequently used the SRP-CRISP-DM framework in practice

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied computing and informatics*, 15(1), 27-33. https://doi.org/10.1016/j.aci.2017.09.005.

A Machine Learning Framework for Sport Result Prediction

Rory P. Bunker[1], Fadi Fayez[2]

[1]Auckland University of Technology, Auckland, New Zealand
[2] Nelson Marlborough Institute of Technology, Auckland, New Zealand

Abstract

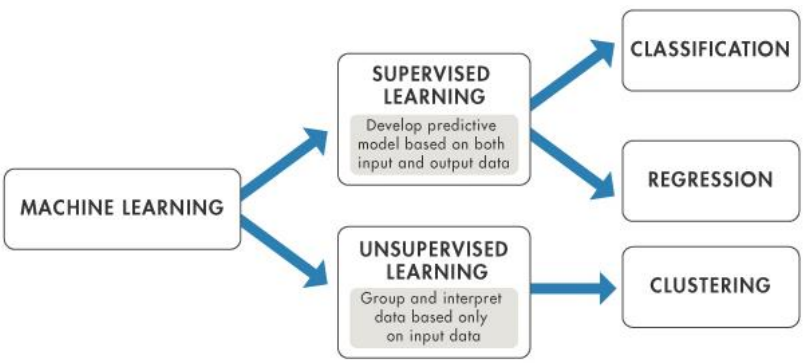
Machine learning (ML) is one of the intelligent methodologies that have shown promising results in the domains of classification and prediction. One of the expanding areas necessitating good predictive accuracy is sport prediction, due to the large monetary amounts involved in betting. In addition, club managers and owners are striving for classification models so that they can understand and formulate strategies needed to win matches. These models are based on numerous factors involved in the games, such as the results of historical matches, player performance indicators, and opposition information. This paper provides a critical analysis of the literature in ML, focusing on the application of Artificial Neural Network (ANN) to sport results prediction. In doing so, we identify the learning methodologies utilised, data sources, appropriate means of model evaluation, and specific challenges of predicting sport results. This then leads us to propose a novel sport prediction framework through which ML can be used as a learning strategy. Our research will hopefully be informative and of use to those performing future research in this application area.

Keywords: Machine learning, event forecasting, data mining, sport result prediction

1. Introduction

One of the common machine learning (ML) tasks, which involves predicting a target variable in previously unseen data, is classification (Mohammad, et al., 2015). The aim of classification is to predict a target variable (class) by building a classification model based on a training dataset, and then utilising that model to predict the value of the class of test data (Witten, et al., 2011). This type of data processing is called supervised learning since the data processing phase is guided toward the class variable while building the model. Some common applications for classification include loan approval, medical diagnoses, email filtering, among others (Abdelhamid & Thabtah, 2014).

Figure 1. Supervised Learning versus Unsupervised Learning (Mathworks, n.d.)



Sport prediction is usually treated as a classification problem, with one class (win, lose, or draw) to be predicted (Prasitio & Harlili, 2016). Although some researchers (e.g. Delen, et al., 2012), have also looked at the numeric prediction problem, where they predict the winning margin – a numeric value. In sport prediction, large numbers of features can be collected including the historical performance of the teams, results of matches, and data on players, to help different stakeholders understand the odds of winning or losing forthcoming matches. The decision of which team is likely to win is important because of the financial assets involved in the betting process; thus bookmakers, fans, and potential bidders are all interested in approximating the odds of a game in advance (Fernandez & Ulmer, 2014). Once a predicted result for the match is obtained, an additional problem is to then decide whether to bet on the match, given the bookmaker's odds. In addition, sport managers are striving to model appropriate strategies that can work well for assessing the potential opponent in a match (Nunes & Sousa, 2006). Therefore, the challenge of predicting sport results is something that has long been of interest to different stakeholders, including the media. The increasing amount of data related to sports that is now electronically (and often publically) available, has meant that there has been an increasing interest in developing intelligent models and prediction systems to forecast the results of matches.

In this paper, we provide a critical survey of the literature on ML for sport result prediction, focusing on the use of neural network (NN) for this problem. Several studies in the statistical and operations research literature have previously considered sport results prediction, but the use of the NN paradigm for this purpose is a more recent area of study. The powerful NN technique has proven to be effective in deriving highly accurate classification models in other domains (Mohammad, et al., 2016). Discussions on the challenges that arise when using these intelligent models for sport results prediction is also provided. Our main contribution is that a CRISP-DM type framework for sport result prediction is proposed (SRP-CRISP-DM), based on the six steps of the standard CRISP-DM framework (Shearer, 2000). This paper serves researchers, sport fans, club managers, bookmakers, academics, and students who are interested in intelligent solutions based on NN for the challenging problem of sport results prediction. This paper will be of use to those who are interested in pursuing future research within this application domain.

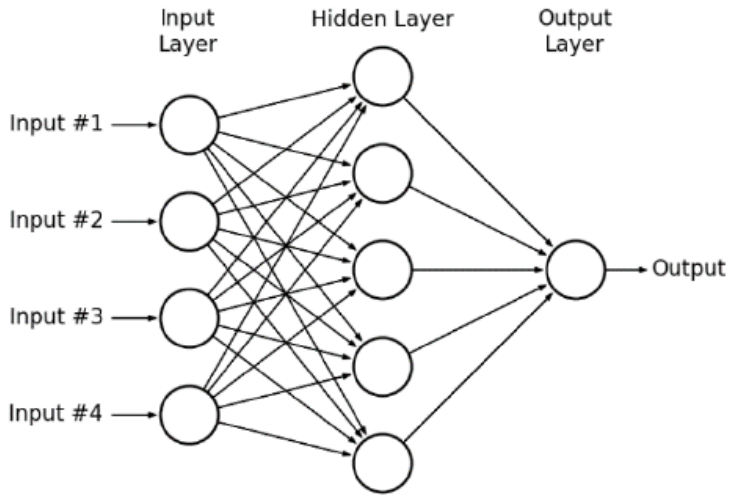
This remainder of this paper is organised as follows. In section 2, studies that have used ANN exclusively, which was the key approach used in earlier research papers in the sport prediction application, are reviewed. Section 3 then provides critical discussion and observations on prior work in this application domain, in the context of the proposed SRP-CRISP-DM framework, conventional measures of model performance, and how we propose that model performance should be measured for the problem of sport results prediction. Finally, section 4 concludes the paper.

2. Literature Review and Critical Analysis

Artificial Neural Networks (ANNs) (Grossberg, 1988) are perhaps the most commonly applied approach among ML mechanisms to the sport result prediction problem. Thus, for this review, we focus on studies that have applied ANNs. An ANN usually contains interconnected components (neurons) that transform a set of inputs into a desired output (Witten, et al., 2011). See figure 2 for an example of an ANN structure. The power of ANN comes from the non-linearity of the hidden neurons in adjusting weights that contribute to the final decision. ANN output often relies on input features and other components associated with the network, such as these weights. The ANN model is constructed after processing the training dataset that contains the features used to build the ANN classification model. In other words, weights associated with interconnected components are continuously changing to accomplish high levels of predictive accuracy. These changes are performed by the ANN algorithm to fulfill the desired model's accuracy given earlier by the user. This may lead in some cases to the problem of overfitting, as well as wasting computing resources such as training

time and memory (Mohammad, et al., 2014). An appealing feature of ANNs is that they are quite flexible in terms of how the class variable is defined e.g. whether it is probability of victory (e.g. McCabe & Travanthan, 2008), or whether two classes are used e.g. with home goals and away goals represented in the two different classes (e.g. Arabzad et al., 2014).

Figure 2. Example structure of an ANN with 4 input nodes in the input layer, 5 hidden nodes in the hidden layer and one output node in the output layer (Mohamed et al., 2015, p.252).



Purucker (1996) conducted one of the initial studies on predicting results in the National Football League (NFL) using an ANN model. Data from the first eight rounds of the competition and five features were used, consisting of yards gained, rushing yards gained, turnover margin, time of possession, and betting line odds. Unsupervised methods based on clustering were used to distinguish between good and poor teams. An ANN with backward-propagation (BP) was then used (Rumelhart, et al., 1986). Purucker achieved 61% accuracy compared with 72% accuracy of the domain experts. The BP algorithm was found to be the most effective approach. A limitation of this study is that only a relatively small number of features were used.

Kahn (2003) extended the work of Purucker (1996) and achieved greater accuracy, performing slightly better than experts in the NFL who were making predictions on the same games. Data on 208 matches in the 2003 season were collected. The features that were used were: total yardage differential, rushing yardage differential, turnover differential, away team indicator and home team indicator. There were two classes: away team outcome and home team outcome – a value of -1 indicating that the team lost the match, and a value of +1 indicating that the team won the match. The problem was treated as a classification problem. The first 192 matches were used as the training data set, and the remaining rounds (week 14 and 15) were used as the test set. Through testing, a network structure of 10-3-2 was found to be optimal. Accuracy of 75% was achieved across the week 14 and 15 matches. The results were compared to the predictions of eight sportscasters from ESPN.com. Across the same matches, the domain experts predicted an average of 63% of matches correctly.

McCabe & Trevathan (2008) attempted to predict results in four different sports: NFL (Rugby League), AFL (Australian Rules football), Super Rugby (Rugby Union), and English Premier League Football (EPL) using data back to the year 2002. A multi-layer perceptron, trained with BP and conjugative-gradient algorithms was used. The ANN had 20 nodes in the input layer, 10 nodes in the hidden layer, and 1 node in the output layer (20-10-1). Features that were used were the same across all the sports and attributes related to specific events within a rugby or soccer match were not

considered. The average performance of the ANN algorithm in predicting results was around 67.5%, compared with expert tipster predictions that achieved around 60% to 65% accuracy.

ANN has also been applied by Davoodi & Khanteymoori (2010) to predict the results of horse races. The authors used data from 100 races at the Aqueduct Race Track held in New York during January of 2010. One ANN was used for each horse in the race, with the output being the finishing time of that horse. Eight features were used for the input nodes in each NN. These were horse weight, type of race, horse trainer, horse jockey, number of horses in race, race distance, track condition, and weather. This optimal network architecture (8-2-1), in terms of mean-squared error, consisted of four layers: an input layer (with eight input nodes), two hidden layers, and an output layer (with horse finishing time). Five different training algorithms were applied to the data: gradient-descent BP (BP), gradient-descent with a momentum parameter (BPM), Levenberg-Marquadt (LM), and conjugate gradient descent (CGD) (Rumelhart, et al., 1986; Sutton, 1986; Kanzow, et al., 2004; Yvan, 2000). It was found that with 400 epochs, the BPM (with momentum parameter of 0.7) and the BP algorithms were most effective at predicting the winner of the race, with BP obtaining an accuracy of 77%. However, the disadvantage of BP was that the training time was lengthy (LM had the shortest training time).

Tax and Joustra (2015) used Dutch football competition data from the past 13 years to predict the results of football matches. The authors were interested in how a model with betting odds alone compared with a hybrid model of both betting odds and other match features. Importantly, and something that has most often been missed in previous studies, they mentioned that cross validation is not appropriate for sport prediction because of the time-ordered nature of the data. A structured literature review from statistical and sport science papers was conducted to identify relevant features to include. Principal component analysis (PCA), sequential forward selection, ReliefF attribute evaluation, and correlation based feature subset selection were used (Jolliffe, 2002; Marcano-Cedeno, et al., 2010; Kononenko, et al., 1997; Hall, 1999). Nine classification algorithms were used in the experimentation, utilizing the machine learning software WEKA, namely naive Bayes, LogitBoost (with decision stumps), NN with BP, Random Forest, CHIRP, FURIA, DTNB, C4.5, and hyper pipes (Hall, et al., 2009; Wilkinson, et al., 2011; Hühn, & Hüllermeier, 2009; Hall & Frank, 2008; Quinlan, 1993). The highest performing classifiers on the full feature set were naive Bayes (used with a 3-component PCA), and the ANN (used with a 3 or 7-component PCA). Both achieved a classification accuracy of 54.7%. In a model including only betting odds features, the highest accuracy of 55.3% was achieved with the FURIA classifier, and was slightly higher than the model with the full feature set (although not statistically significant). In a hybrid model of the public data features with the betting odds features, LogitBoost with ReliefF attribute selection provided the highest classification accuracy of 56.1%. The difference between the public data model and the betting odds model was, however, not statistically significant according to McNemar's test. However, this did highlight that betting odds alone can be a reasonable predictor of match outcome.

In non-team sports, researchers have used machine learning models to predict the performance of the individual player.

Maszczyk et al. (2014) compared neural networks and non-linear regression to predict the distance of Javelin throws. The aim of the investigation was to identify the usefulness of neural networks as an athlete recruitment tool, and how this compared to the commonly used regression models. The data set consisted of 70 javelin throws - a training set consisting of 40 cases, a validation set consisting of 15 cases, and a test set consisting of 15 cases. Their initial statistical analysis using a correlation matrix and regression analysis found four significant predictors of Javelin throw length: cross step, specific power of the arms and the trunk, specific power of the abdominal muscles, and grip power. The numeric class variable used was the average distance of three throws from a full run-up after a 30 minute warm up. Through experimentation, the best architecture in terms of normalized root mean squared error, of the neural network was found to be 4-3-1 (four input neurons/variables, one hidden layer with three neurons, and one outcome). The javelin throws of 20 javelin throwers from the Polish national team

were predicted using the models, and were compared with the actual length of the throws. Their results showed that the neural network models offered much higher quality of prediction than the nonlinear regression model. The absolute network error was found to be 16.77 m, versus the absolute regression error of 29.45 m.

Edelmann-Nusser et al (2002) investigated modeling the performance of an elite female swimmer in the finals of the 200 m backstroke at the Olympic Games in 2000 in Sydney. Data consisted of the performance output of 19 competitions in 200 m backstroke prior to the Olympics and data from the swimmer's training period - the last 4 weeks prior to the competition. An MLP with 10 input neurons, 2 hidden neurons and 1 output neuron was used. The results show that the MLP was accurate; the error of the prediction was only 0.05 s. The MLP was also compared with linear regression, which did not provide as accurate results. This paper (as well as Maszcyk et al. (2014)) highlights the potential usefulness for machine learning techniques to be used by high performance staff and analysts in professional sport for identifying the factors to focus on when developing training programs, not just purely for result prediction.

Wiseman (2016) predicted winning PGA golf score based on scores after round 1 of a competition. Note that they were predicting winning score, not tournament winner itself. The authors compared the performance of: linear regression, neural network regression, Bayesian linear regression, decision forest regression and boosted decision tree regression, in the Microsoft Azure service. The authors performed correlation matrix analysis of different features and selected Round 1 leading score, round 1 average score, course par, major event, course yardage and total prizemoney as the predictors. R-squared value and MSE were used to evaluate algorithm accuracy. Data from 2004 to 2015 was used to construct the models, and tournaments from 2016 were used to validate them. Linear regression and Bayesian linear regression were the best performing models on the 2016 data set, predicting the winning score to within 3 shots 67% of the time.

3. The Proposed Sport Result Prediction Intelligent Framework

We would argue that the use of a structured experimental approach to the problem of sport results prediction is useful to obtain the best possible results with a given data set. In this section, an intelligent architecture for sport results prediction is presented, proposing steps of a possible ML framework, and describing the characteristics of the data used for sport results prediction, and how this fits within the framework. Our framework (figure 4) focuses on result prediction for team sports rather than individual sport. Our Sport Result Prediction CRISP-DM framework or SRP-CRISP-DM framework consists of six main steps, based on the steps of the standard CRISP-DM (figure 3) framework (Shearer, 2000).

Figure 3. The six phases of the traditional CRISP-DM Model (Shearer 2000, p.14)

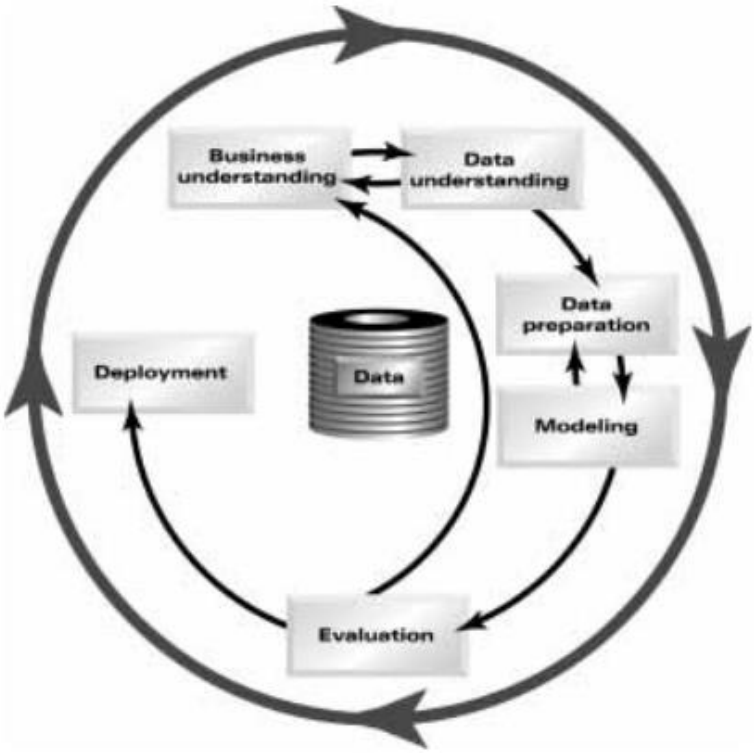
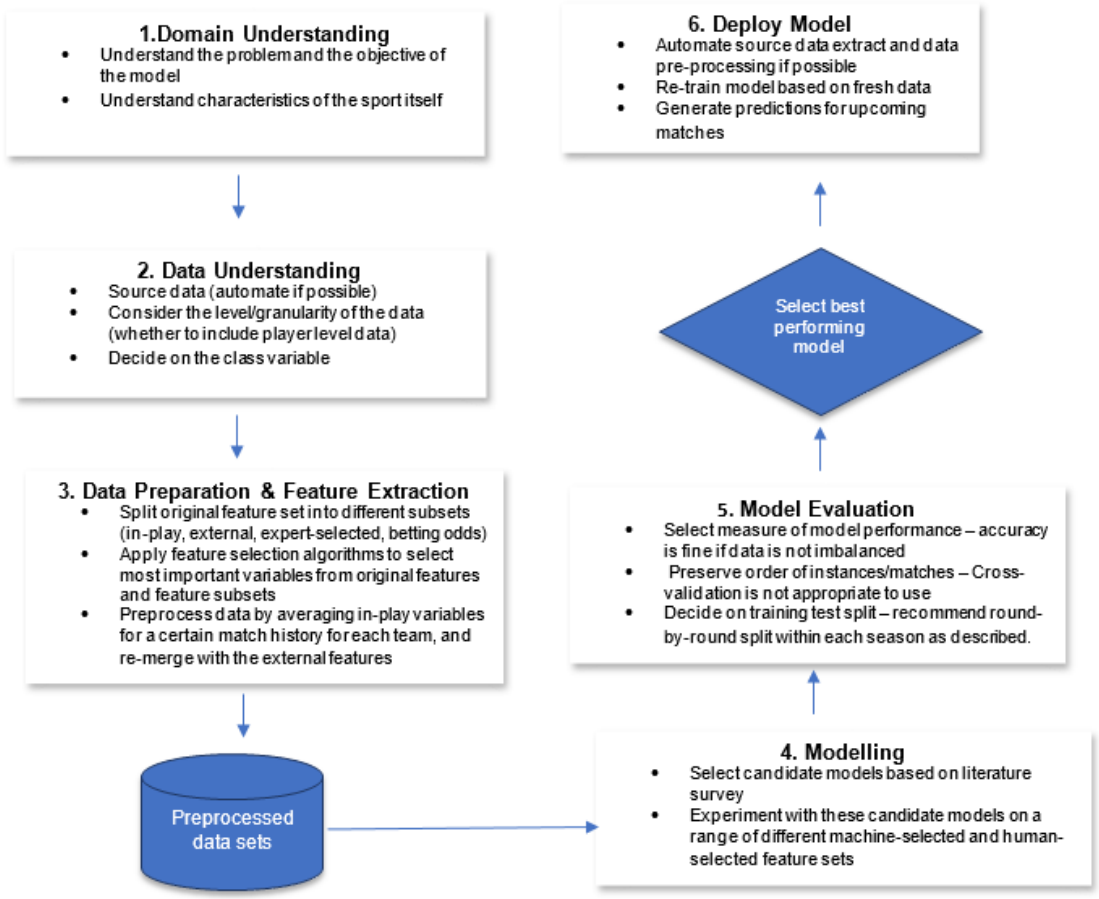


Figure 4. Steps of our proposed SRP-CRISP-DM framework



3.1. Domain Understanding

Domain understanding includes comprehending the problem, the goal of the modeling, and the specific characteristics of the sport itself. This involves having some understanding of how the sport is played and what factors are potentially involved in determining the outcome of matches. This could be obtained through personal knowledge of the sport or obtained by surveying existing literature or by consulting experts in the sport.

There also needs to be clarity regarding the objective of the model. It could be to predict results to compete with expert predictions, online competitions, or it could be to ultimately use the results of the model to bet on matches. If the predictive model is used for betting, there needs to also be consideration of which matches will be bet on. For example, there will likely be some betting odds threshold where, although the model predicts a victory for that team, the betting odds are so low that the return does not warrant betting on that match at all (e.g. if a team is paying $1.01 to win, meaning that a $100 bet placed would only return $1).

3.2. Data Understanding

Data for sport prediction is often able to be obtained online from publically available sources. Some prior studies have automated the data collection process, writing scripts that automatically extract the online data and then load it into some form of database. Some studies have also built an end-user interface, where users can input data for an upcoming match and the prediction is then generated.

The granularity/level of the data is something that needs to be considered. Previous studies have generally had training data that is at the match/team level. It is also possible to include player-level data, which contains statistics on the players that have played in each of the matches. Player level data will generally be contained in a separate data set that would then have to be transposed and joined with the match level data so that each match has certain player statistics as attributes in the data set. Including player level data would have the advantage that we can investigate whether specific players' actions or presence are important for the performance of the team in terms of whether they win or lose.

The definition of the class variable needs to also be considered. Most prior work has treated the sport prediction problem as a 2 or 3 class values classification problem (home win, away win) or (home win, draw, away win). Delen, et al. (2012), also considered the problem as a numeric prediction problem, using regression techniques to predict the points margin (home points minus away points), and then making a win-loss prediction based on the predicted points margin. The authors ultimately found that treating the problem as a classification resulted in superior results, but that is not to say that this would be the case on all sports or all data sets.

3.3. Data Preparation & Feature Extraction

3.3.1. Creating Feature Subsets

Features in sport result data can be divided into several different subsets. Miljković, et al. (2010), for example, split the features into match-related and standings features. Tax and Joustra (2015), considered how a hybrid model of betting odds and public data features compared with a feature set of betting odds alone. Hucaljuk and Rakipović (2011) used a separate expert-selected feature set against their own feature set, to investigate the value of expert opinion for feature selection. And of course, there are feature sets that are selected based on feature selection algorithms – either on the full feature set or a subset of the original feature set. Ideally, researchers should test several different feature selection approaches in conjunction with testing their candidate classification models. Then, as Tax and Joustra (2015) did, one can investigate which classifier and feature selection algorithm together produces the best classification accuracy.

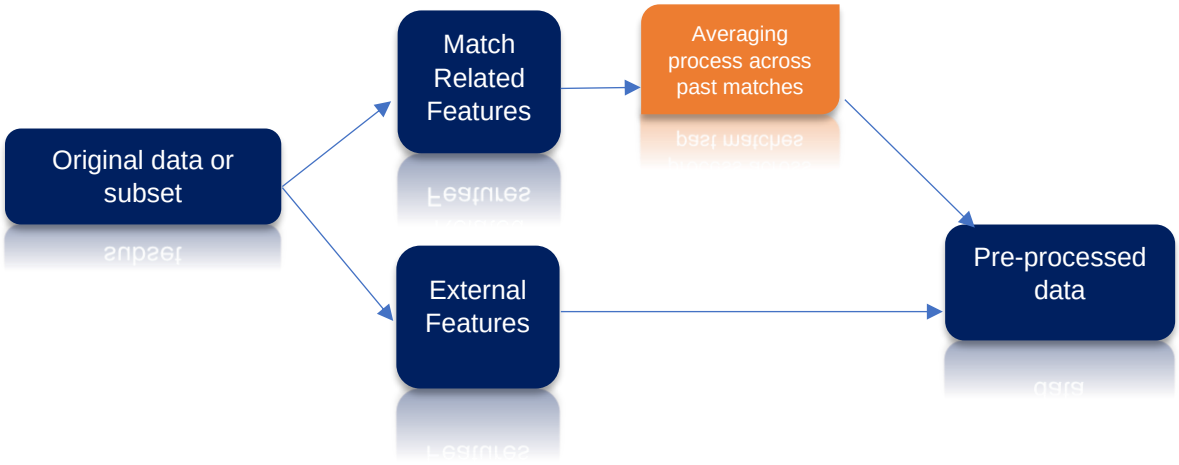
Hucaljuk and Rakipović (2011) included an expert-selected feature set in addition to their initial feature set. Although this was not found to result in improved accuracy, this could depend on the characteristics of the sport, and perhaps the experts themselves. Another way that expert opinion can be used is in comparing the predictive accuracy of the predictive models, with the predictions of the experts. To incorporate expert opinion, one could either generate an expert-selected feature set to compare with machine-learned feature selection approaches, or alternatively, compare their model with expert predictions.

3.3.2. Data Preprocessing: Match Features versus External Features

There can be a distinction made between 'match-related' and 'external' features. Match-related features relate to actual events within the sport's match. For example, in football these could be meters gained, passes made, and so on. External features do not relate to events within the match, that is are external to the match itself (e.g. recent form, travel, players available for the match, etc.). This distinction is important for data preprocessing purposes. External features are known prior to the upcoming match to be played. For example, we know the distance that both teams have travelled and

we know both teams' recent form leading into the upcoming match. Match-related features however, are not known until the match has been played. Thus, we only know an average of these features for a certain number of past matches for these teams. For example, we would know the average passes made per match by both teams prior to the match, but do not know the actual passes made in the upcoming match until after it has been played. This means that only past average statistics for these features can be used to predict an upcoming match. Therefore, match-related features should undergo a separate averaging process before being re-merged with the external features. Buursma (2011) followed this process, and found, through experimentation, that using an average across the past 20 matches resulted in the best classification accuracy.

Figure 5. Match-Related statistics should go through an averaging process across a certain number of historical matches for each team, and then be re-merged with the external match features.



3.4. Modelling

The first step in the modelling process is to select which candidate models will be used in the experimentation. This would involve a review of past literature, and identification of commonly applied predictive models that have previously been successful. Each model can then be trialed on each feature subset, and subsets that have been selected by feature selection algorithms. Experimentation with these different feature selection methods and classification models will identify the best combination of classifier and feature selection technique.

3.5. Sport Prediction Model Evaluation

3.5.1. Measuring Model Performance

To evaluate model performance, one would classify match results into home wins, away wins and draws (if the sport has draws) and then look at the number of matches that the model has correctly identified, using a standard classification matrix. There is unlikely to be a great degree of imbalance in the class values for the dataset, although given the commonly observed home advantage phenomenon, one is likely to see a slight skew in favor of home wins. In this case, classification accuracy is a reasonable measure of evaluation. In cases where the data is highly imbalanced, ROC curve evaluation may be more appropriate.

3.5.2. Training and Testing

As has been mentioned, it is important to preserve the order of the training data for the sport prediction problem, so that upcoming matches are predicted based on past matches only. Cross-validation (figure 6) generally involves shuffling the order of the instances and therefore is not an appropriate means of splitting the data into training and testing, for the sport result prediction problem. A held-out training test split is more appropriate, with the order of the instances being preserved. Machine learning software such as WEKA provide the option to preserve the order of instances.

Figure 6. Diagrammatic representation of 10-fold Cross-Validation (Raschka, n.d.)

An appropriate training-test split needs to be decided on. This may depend on the amount of data that the researcher has on hand – whether they have only one season of data, or multiple seasons. Usually professional sport competitions are organized in rounds, with teams playing matches over the weekend. Teams usually play one match in each round unless they have a 'bye'. In the case where one season of data is on hand, the number of rounds that will be used for training the model, and the number of rounds that will be used for testing the model needs to be determined. For example, in a data set with 10 rounds of data, the first 7 rounds of the competition could be used for training the model and the last 3 rounds of the competition could be used for testing the model. However, to obtain a more realistic measure of model performance, round 1 could be used as training to test on round 2, round 1 & 2 could be used as training to test on round 3, round 1 – 3 could then be used as training to test on round 4, and so on. So, within a season which contains a certain number of competition rounds, we use rounds 1 to n-1 to train our model, and use round n as the test data set, for each round n in N, where N is the total number of rounds in the competition. We thus obtain a classification accuracy for each of these training/test splits, and take an average of the accuracies to give an overall measure of model performance.

Rather than round-by-round prediction, another possibility is to update the training data set after every match has been played. In this case, all past matches up to the current match as training data, and the upcoming match as the training data (i.e. only having that one record as the training data). This is essentially like order-preserved leave-one-out cross-validation. This match-by-match approach is probably not necessary unless teams play more than one match over the same competition round.

Some papers have used multiple seasons of data. A common approach has been to use earlier seasons as training data, to predict the later seasons as the test data set. For example, Cao (2012) used seasons up from 2005/2006 to 2009/2010 were used as the training data, and the 2010/2011 season was used as the test data, to evaluate models that predict basketball match results. Prior seasons may not be relevant to predict matches in future seasons, particularly in sports where team rosters and strengths can change

significantly from year to year. This approach may not give a reliable picture of model performance (although this could be mitigated to some extent if player level data is included, and so player changes would be captured from season to season). We would argue that, although more computationally intensive and laborious, our round-by-round training test split approach mentioned above should be used *within each season*. An average model classification accuracy could then be produced for each season, and a plot could be shown of model accuracy by season.

3.6. Model Deployment

Ideally, one can automate the process so that new round data is obtained from the web, and added to the match database (or otherwise added to the database manually by the end-user). The training data and test data are then adjusted, the model is retrained with the new training data, and new matches are predicted. Predictions are then returned to the end user. The learning model in the proposed architecture could also be online and dynamically receiving input data prior to the match beginning (external features) and while the match is played (match features). It also should be incremental in the way that the training data set is continuously updated, and thus the classifier would keep changing to reflect those of the learning environment.

4. Conclusions

One of the vital applications in sport that requires good predictive accuracy is match result prediction. Traditionally, the results of the matches are predicted using mathematical and statistical models that are often verified by a domain expert. Due to the specific nature of match-related features to different sports, results across different studies in this application can generally not be compared directly.

Despite the increasing use of ML models for sport prediction, more accurate models are needed. This is due to the high volumes of betting on sport, and for sport managers seeking useful knowledge for modelling future matching strategies. Therefore, ML seems an appropriate methodology for sport prediction since it generates predictive models that can predict match results using predefined features in a historical dataset.

This article critically analyses some recent research on sport prediction that have used ANN, and following this, we proposed a sport result prediction ‘SRP-CRISP-DM’ framework for the complex problem of sport result prediction. Moreover, challenges facing the sport prediction application were shown to pinpoint future work for scholars in this important application. Future studies concerning ML in sport result prediction research will hopefully be benefitted by this study.

References

1. Abdelhamid N., and Thabtah F. (2014). Associative Classification Approaches: Review and Comparison. *Journal of Information and Knowledge Management (JIKM)*, 13(3).
2. Agrawal R., and Srikant R. (1994). Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.
3. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46 (3): 175–185.
4. Arabzad, A. C., Araghi, M. E. T., & Soheil, S. N. (2014). Football Match Results Prediction Using Artificial Neural Networks: The Case of Iran Pro League. *International Journal of Applied Research on Industrial Engineering*, *1*(3), 159-179.
5. Bouckaert, R. R. (2004). Bayesian network classifiers in Weka. (Working paper series. University of Waikato, Department of Computer Science. No. 14/2004). Hamilton, New Zealand: University of Waikato.
6. Cao, C. (2012). Sports data mining technology used in basketball outcome prediction. Master's Thesis, Dublin Institute of Technology, Ireland.
7. Cortes C. and Vapnik V. (1995). Support-Vector Networks. *Machine Learning*, 20(3): 273 – 297.
8. Davoodi, E., and Khanteymoori, A. (2010). Horse racing prediction using artificial neural networks. *Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing*, 2010: 155–160.
9. Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting*, *28*(2), 543-552.
10. Duda, R. O. and Hart P. E. (1973). *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons.
11. Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2002). Modeling and prediction of competitive performance in swimming upon neural networks. *European Journal of Sport Science*, *2*(2), 1-10.
12. Fernandez, M., and Ulmer, B. (2014). Predicting Soccer Match Results in the English Premier League.
13. Friedman J., Hastie T., and Tibshirani R. (2000) Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28(2): 337–407.
14. Grossberg S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, 17–61.
15. Hall, M. A. (1999). Correlation-based Feature Subset Selection for Machine Learning. PhD dissertation, Department of Computer Science, University of Waikato.
16. Hall, M. and Frank, E. (2008). Combining naive Bayes and decision tables. Proc 21st Florida Artificial Intelligence Research Society Conference, Miami, Florida. AAAI Press.
17. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., and Witten I. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10-18.
18. Hühn, J., & Hüllermeier, E. (2009). FURIA: an algorithm for unordered fuzzy rule induction. *Data Mining and Knowledge Discovery*, *19*(3), 293-319.
19. Hofmann, M., and Klinkenberg, R. (Eds.). (2013). RapidMiner: Data mining use cases and business analytics applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, CRC Press.
20. Hucaljuk J., and Rakipović A. (2011). Predicting football scores using machine learning techniques. In *MIPRO, 2011 Proceedings of the 34th International Convention, IEEE*, 1623–1627.
21. Jolliffe, I. T. (2002). *Principal Component Analysis*, second edition. Berlin: Springer-Verlag.

22. Joseph A., Fenton N.E., and Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19:544–553.
23. Kahn, J. (2003). Neural network prediction of NFL football games. *World Wide Web electronic publication* (2003), 9–15.
24. Kanzow C., Yamashita N., and Fukushima M. (2004). Levenberg-Marquardt methods with strong local convergence properties for solving nonlinear equations with convex constraints. *JCAM*, 172(2):375-97.
25. Kononenko, Igor, et al. (1997). Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, 7(1): 39-55.
26. Marcano-Cedeno A, Quintanilla-Dominguez J, Cortina-Januchs M, and Andina D (2010). Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In: 36th annual conference on IEEE industrial electronics society, pp 2845–2850
27. Maszczyk, A., Gołaś, A., Pietraszewski, P., Roczniok, R., Zając, A., & Stanula, A. (2014). Application of neural and regression models in sports results prediction. Procedia-Social and Behavioral Sciences, 117, 482-487.
28. Mathworks (n.d.). machinelearning_supervisedunsupervised.png. Retrieved from https://de.mathworks.com/help/stats/machinelearning_supervisedunsupervised.png
29. McCabe A. and Trevathan J. (2008). Artificial intelligence in sports prediction. In *Information Technology: New Generations*, ITNG 2008. Fifth International Conference on, IEEE, 1194–1197.
30. Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010). The use of data mining for basketball matches outcomes prediction. In *Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on* (pp. 309-312). IEEE.
31. Mohammad R., Thabtah F., and McCluskey L. (2015). Tutorial and critical analysis of phishing websites methods. *Computer Science Review Journal*, 17: 1–24.
32. Mohamed, H., Negm, A., Zahran, M., & Saavedra, O. C. (2015). Assessment of artificial neural network for bathymetry estimation using High Resolution Satellite imagery in Shallow Lakes: case study El Burullus Lake. In *International water technology conference*.
33. Nunes, S., and Sousa, N. (2006). Applying Data Mining Techniques to Football Data from European Championships. *Proceedings of Conferência de Metodologias de Investigação Científica*, 4-16.
34. Purucker M. C. (1996). Neural network quarterbacking. *IEEE Potentials*, 15: 9–15.
35. Prasitio D., and Harlili D. (2016). Predicting football match results with logistic regression. Proceedings of the 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA). 16-19 Aug. 2016, Penang, Malaysia.
36. Quinlan, J. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.
37. Raschka (n.d.). k-fold.png. Retrieved from https://sebastianraschka.com/images/faq/evaluate-a-model/k-fold.png
38. Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323 (6088): 533–536.
39. Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, *5*(4), 13-22.
40. Sutton, R. S. (1986). Two problems with backpropagation and other steepest-descent learning procedures for networks. Proc. 8th Annual Conf. Cognitive Science Society
41. Thabtah F., Hammoud S. and Abdeljaber H. (2015). Parallel Associative Classification Data Mining Frameworks Based Mapreduce. To Appear in *Journal of Parallel Processing Letter*. March 2015. World Scientific.
42. Thabtah F., Mohammad R. M., McCluskey L. (2016). A dynamic self-structuring neural network model to combat phishing. Neural Networks (IJCNN), 2016 International Joint Conference, pp. 4221-4226. Vancouver, Canada, 2016.

43. Wilkinson, L., Anand, A., and Tuan, D. (2011). CHIRP: a new classifier based on composite hypercubes on iterated random projections. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 11:6–14.
44. Wiseman, O. (2016). *Using Machine Learning to Predict the Winning Score of Professional Golf Events on the PGA Tour* (Doctoral dissertation, Dublin, National College of Ireland).
45. Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann.
46. Yvan, N. (2000). Flexible Conjugate Gradients. *SIAM Journal on Scientific Computing*, 22 (4): 1444.

CHAPTER 5: PUBLICATION 2 - “SUPERVISED SEQUENTIAL PATTERN MINING OF EVENT SEQUENCES IN SPORT TO IDENTIFY IMPORTANT PATTERNS OF PLAY: AN APPLICATION TO RUGBY UNION”

This paper investigated performance at the sequence of play level, utilising data consisting of event sequences derived from event log data labelled with a binary score/did not score label. This study employed a discriminative sequential pattern mining technique called the S3P-classifier (Nakagawa et al., 2016; Sakuma et al., 2019). As the name suggests, this method is a discriminative sequential pattern mining technique that is also a machine learning classifier. The S3P-classifier incorporates the safe pattern pruning (SPP) algorithm proposed by Nakagawa et al. (2016) and Sakuma et al. (2019) to efficiently prune patterns (in this case, subsequences of events) that are guaranteed not to discriminate between the sequence labels.

The S3P-classifier was applied to an event log dataset containing 490 "passage of play" sequences from one team's matches in the 2018 Japan Top League Rugby season. These were labelled based on whether they were ultimately scoring or non-scoring. The event subsequences that discriminated most between scoring and non-scoring outcomes, from both the team's and opposition teams' perspectives, were obtained using the S3P-classifier. These key patterns were then compared with the most frequent patterns that were obtained with well-known unsupervised sequential pattern mining algorithms (specifically, PrefixSpan (Pei et al., 2004), the Generalized Sequential Pattern (GSP) algorithm (Srikant & Agrawal, 1996), Fast (Salvemini, Fumarola, Malerba, & Han, 2011), and CM-SPADE and CM-SPAM (Fournier-Viger et al., 2014)) when applied to subsets of the same dataset that were partitioned on the label (this partitioning means that unsupervised methods possessed knowledge of which sequences were scoring).

Regarding the practical implications of the study, the event subsequence patterns obtained by the S3P-classifier were readily interpretable. It was found that line breaks, successful lineouts, regained kicks in the field of play, repeated phase-breakdown play, and failed exit plays by the opposition team were the key patterns that discriminated most between the team scoring and not scoring. Opposition team line breaks, errors made by the team, opposition team lineouts, and repeated phase-breakdown play by the opposition team were

the patterns that discriminated most between the opposition team scoring and not scoring. Notably, compared to the patterns obtained by the unsupervised sequential pattern mining methods, the patterns obtained by the supervised S3P-classifier method contained a much greater variety of events for interpretation by coaches and analysts for performance analysis purposes.

This work's significance in terms of its contribution to the literature was in the application of a supervised sequential pattern mining technique that had not previously been applied in the context of sports. The study found that discriminative sequential pattern mining methods such as the S3P-classifier can be a valuable tool for analysing event sequences in sports by identifying key patterns of events that discriminate between successful and unsuccessful outcomes. As mentioned, the discriminative sequential pattern mining technique was also found to yield much more useful event subsequences for interpretation by coaches.

Bunker, R., Fujii, K., Hanada, H., & Takeuchi, I. (2021). Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: an application to rugby union. *PloS one*, 16(9), e0256329. https://doi.org/10.1371/journal.pone.0256329.

Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: an application to rugby union

Rory Bunker[1*], Keisuke Fujii[1,2], Hiroyuki Hanada[2], Ichiro Takeuchi[2,3]

1 Graduate School of Informatics, Nagoya University, Nagoya, Aichi, Japan
2 RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
3 Department of Computer Science, Nagoya Institute of Technology, Nagoya, Aichi, Japan

* Corresponding author
Email: rorybunker@gmail.com

Abstract

Given a set of sequences comprised of time-ordered events, sequential pattern mining is useful to identify frequent subsequences from different sequences or within the same sequence. However, in sport, these techniques cannot determine the importance of particular patterns of play to good or bad outcomes, which is often of greater interest to coaches and performance analysts. In this study, we apply a recently proposed supervised sequential pattern mining algorithm called safe pattern pruning (SPP) to 490 labelled event sequences representing passages of play from one rugby team's matches in the 2018 Japan Top League season. We obtain patterns that are the most discriminative between scoring and non-scoring outcomes from both the team's and opposition teams' perspectives using SPP, and compare these with the most frequent patterns obtained with well-known unsupervised sequential pattern mining algorithms when applied to subsets of the original dataset, split on the label. From our obtained results, line breaks, successful line-outs, regained kicks in play, repeated phase-breakdown play, and failed exit plays by the opposition team were found to be the patterns that discriminated most between the team scoring and not scoring. Opposition team line breaks, errors made by the team, opposition team line-outs, and repeated phase-breakdown play by the opposition team were found to be the patterns that discriminated most between the opposition team scoring and not scoring. It was also found that, probably because of the supervised nature and pruning/safe-screening mechanisms of SPP, compared to the patterns obtained by the unsupervised methods, those obtained by SPP were more sophisticated in terms of containing a greater variety of events, and when interpreted, the SPP-obtained patterns would also be more useful for coaches and performance analysts.

Introduction

Large amounts of data are now being captured in sport as a result of the increased use of GPS tracking, optical, and video analysis systems, as well as enhancements in computing power and storage. There is great interest in making use of this data for performance analysis purposes. A wide variety of methods have been used to analyze sports data, ranging from statistical methods to, more recently, machine learning, deep learning and data mining techniques.

Among the various analytical frameworks available in sports analytics, in this paper, we adopt an approach to extract events from sports matches and analyze sequences of events. The most basic events-based approach is based on the analysis of the frequencies of events, which can be used as performance indicators [1]. Alternatively, by comparing the frequency of each event in sequences with positive outcomes (winning, scoring points, etc.) with the frequency of each event in sequences with negative outcomes (losing, conceding points, etc.), one can investigate which events are commonly associated with these outcomes. However, frequency-based analyses have drawbacks in that the information contained in the order of events cannot be exploited.

In this study, we consider sequences of events, and refer to a partial sequence of events as a sequential pattern, pattern of play, subsequence, or simply a pattern. In sports, the occurrence of certain events in a particular order often has a strong influence on outcomes, so it is useful to use patterns as a basic analytical unit. Invasion sports, e.g., rugby, soccer and basketball, have many events and patterns that occur very frequently and repeatedly. However, although there may be a paucity of events that are important for scoring, these are the patterns that are of greater interest to coaches and performance analysts. For instance, in soccer, a pattern consisting of an accurate cross followed by a header that is on target will occur much less frequently than a pattern consisting of repeated passes between players, but the former pattern is likely to be of much greater interest to coaches and performance analysts because there is a good chance that the pattern may lead to a goal being scored.

The computational framework for finding patterns from sequential data that have specific characteristics is known as sequential pattern mining (SPM) in the field of data mining. The most basic problem setup in SPM is to enumerate frequent patterns, which is called frequent SPM. Although the total number of patterns (i.e., the number of ordered sequences of all possible events) is generally very large, it is possible to efficiently enumerate patterns that appear more than a certain number of times by making effective use of branch-and-bound techniques. In the field of machine learning, frequent SPM is categorised as an unsupervised learning technique.

When applying frequent SPM to data from sport, there are several options. The first option is to simply extract the frequent patterns from the entire dataset. The drawback of this approach is that it is not possible to determine whether a particular pattern leads to good or bad outcomes. The second option is to split the dataset into a "good-outcome" dataset and a "bad-outcome" dataset, and apply frequent SPM to each dataset. The third option is to apply frequent SPM to the entire dataset in order to identify frequent patterns, and then create a machine learning model that uses these patterns as features to predict whether the pattern is associated with good or bad outcomes. A disadvantage of the second and third options is that the pattern extraction process is conducted separately from the process that associates patterns with outcomes.

Unlike unsupervised SPM, supervised SPM directly extracts patterns that are associated with good or bad outcomes. Roughly speaking, by using supervised SPM, we can identify patterns that differ in frequency according to the outcomes in a more direct manner.

Related Work

Sequential pattern mining (SPM)

Sequential pattern mining (SPM) [2] involves discovering frequent subsequences as patterns from a database that consists of ordered event sequences, with or without strict notions of time [3]. Originally used to analyze biological sequences [4–7], SPM methods have also been applied for other purposes including XML document classification [8], keyword/key-phrase extraction [9–11], as well as next item/activity prediction and

recommendation systems [12–17]. For an overview of the SPM field, we refer the reader to [18].

A recent area of interest in SPM has been high-utility SPM, which built on the idea of high-utility pattern (itemset) mining [19] by taking into consideration the utility of patterns. For instance, [20] provides an example where in market basket analysis, although a diamond may sell much less frequently than an egg, it is of much higher utility (profit) and is thus of greater interest to a business. An early study was that of [21], who defined the high-utility SPM problem and proposed a method for it called USpan. Other proposed high-utility SPM methods include HUS-SPAN [22], which was expanded on by [23], who used pruning methods to decrease the required pattern search space in order to improve efficiency in terms of run-time and the number of candidate patterns generated. Scalable high-utlity SPM methods that can be applied to big data (e.g., IoT) have been proposed by [24], who introduced a scalable Spark-based platform, and [25], who proposed a Hadoop-based high fuzzy utility pattern mining (HFUPM) algorithm.

In market basket analysis applications of high-utility SPM, per-item prices or profits can be used as utility values in order to then determine which patterns are interesting; however, in sport, it is difficult to determine explicit *a priori* values for events or patterns. Thus, to overcome this problem, in this study, by assigning outcome labels to sequences based on scoring outcomes and taking a discriminative approach that employs supervised learning, we can instead identify interesting patterns by determining their *a posteriori* values.

One of the first (standard) SPM algorithms to be proposed was GSP [26], which was based on the A-priori algorithm proposed by the same authors [27]. Algorithms including SPADE [28], SPAM [29], and PrefixSpan [30] were later proposed to address some of the limitations that were identified with GSP. PrefixSpan is categorized as a pattern-growth algorithm, since it grows a tree structure of patterns extending from a pattern with a single event at its base, and adds greater numbers of events to the patterns in each of its descendent nodes for all possible patterns in a database. More recently, CM-SPAM and CM-SPADE [31] as well as Fast [32] have been proposed to provide further improvements in computational efficiency and speed compared to the original SPAM and SPADE algorithms. It should be noted that all of the above-mentioned SPM methods are unsupervised methods, which are applied to unlabelled sequences.

Safe pattern pruning (SPP), proposed by [33,34], combines a convex optimisation technique called safe screening [35] with SPM. SPP is supervised, and can be applied to labelled sequences. SPP uses PrefixSpan to grow the initial pattern tree, and redundant patterns are then removed using a specific pruning criterion (see [33,34] for more details on this pruning criterion). In particular, the tree structure grown by PrefixSpan is pruned in such a way that if a node corresponding to a particular pattern is pruned, it is guaranteed that all patterns corresponding to its descendant nodes are not required for the predictive SPP model (Fig 1).

In the SPP method, each pattern is multiplied by a weight, and these weights are calculated by solving an optimization problem, as will be described later in this paper. The magnitude of the weight of a pattern reflects the degree to which that pattern discriminates between positive and negative outcomes (labels). As mentioned, the SPP method also incorporates safe screening, which eliminates redundant weights that are guaranteed to be non-discriminative in the optimal solution (i.e., will have a weight value of zero). SPP has previously been applied to datasets consisting of animal trajectories [34]; however, compared with animal trajectories, data in sport often contains a greater diversity of events.

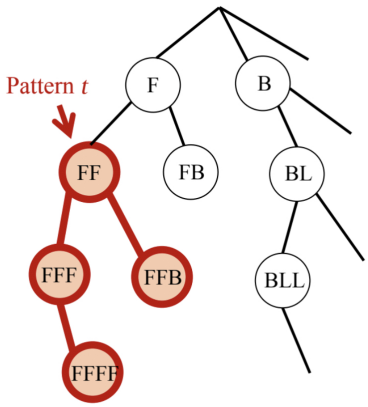


Fig 1. SPP pruning. One of the mechanisms of SPP identifies and removes patterns that do not contribute to the model before performing the optimization. For example, if pattern t does not satisfy the pruning criterion specified in [34], the sub-tree below pattern t is deleted.

Application of SPM methods in sport

Some previous studies have applied unsupervised SPM methods to sport, and these are summarized in Table 1. These prior studies have focused mainly on the identification, interpretation and visualization of sequential patterns. CM-SPAM was applied for performing technical tactical analysis in judo by [36], while sequential data that was obtained using trackers was used to test for significant trends and interesting sequential patterns in a single cyclist's training regime over an extended period of time by [37]. In the context of team sport (soccer), Decroos et al. [38] combined clustering and CM-SPADE using a five-step approach that is summarized in the fifth column of Table 1, and the authors provided a ranking function that allowed the user (e.g., a coach) to assign higher weights to events that are more relevant. For instance, the authors note that, despite their frequency, normal passes are not as of much relevance to coaches as shots and crosses.

Analysis of sequences in rugby union

In the sport of rugby union (hereafter referred to simply as rugby), some prior studies have considered sequences of play by analyzing their duration. For example, [39] studied the durations of sequences of plays leading to tries at the 1995 Rugby World Cup (RWC), and [40] found that, at the 2003 RWC, teams that were able to create movements that lasted longer than 80 seconds were more successful. More recently, [41] applied K-modes cluster analysis to sequences of play in rugby, and found that scrums, line-outs and kick receipts were common scenarios that led to tries being scored in the 2018 Super Rugby season. By employing convolutional and recurrent neural networks, [42] aimed to predict the outcomes of sequences of play (e.g., whether the sequence of play ended with territory gain, retaining of possession, scoring of a try, or conceding/being awarded a penalty), based on the order of events and their on-field locations.

Motivation and Contributions

In this study, we apply SPP, a supervised SPM method, to data consisting of event sequences from all of the matches played by a professional rugby union team in their

Table 1. Prior studies that have applied sequential pattern mining techniques in sport.

Study	Sport	Model Used	Model Type	Summary of Approach	Evaluation Metrics
Hrovat (2015)	Cycling	SPADE	Unsupervised	Applied SPADE to identify frequent sequential patterns, calculated interestingness measures (p-values) for these frequent patterns, and visualized these patterns for increasing/decreasing daily and duration trends	Support, permutation test p-values
La Puma & Giorno (2017)	Judo	CM-SPAM	Unsupervised	Identified patterns with SPM for the tactical analysis of judo techniques	Support
Decroos (2018)	Soccer	CM-SPADE	Unsupervised	Clustered phases based on spatio-temporal components, ranked these clusters, mined the clusters to identify frequent sequential patterns, used a ranking function (weighted support function)—in which a coach can assign higher weights to more relevant events—to score obtained patterns, and then interpreted the obtained patterns	Support (events were weighted according to their relevance based on the judgement of the user), and identified the top-ranked frequent sequences in the clusters

2018 Japan Top League season. The present study is motivated by the fact that, although SPM methods have been applied to sport, only unsupervised SPM methods appear to have been used to date. In addition, no form of SPM method, unsupervised or supervised, appears to have been applied to analyze sequences of play in rugby.

We compare the most discriminative SPP-obtained patterns (subsequences) with the most frequent patterns obtained by well-known unsupervised SPM methods (PrefixSpan, GSP, Fast, CM-SPADE and CM-SPAM), where the unsupervised SPM methods are assumed to possess knowledge of which sequences contain scoring events (i.e., when the unsupervised methods are applied to label-based partitions of the original labelled data).

The main contributions of this study are in the comparison of the usefulness of supervised and unsupervised SPM methods when applied to event sequence data in sport, the application of a supervised SPM method to event sequence data in sport, and the application of a SPM method to analyze sequences of play in rugby.

Notation

Sequences consist of ordered events drawn from a set of m unique event symbols, denoted $\mathcal{S} := \{s_1, \ldots, s_m\}$. Let n denote the number of sequences in the dataset. The sets of sequences with labels 1 and -1 are denoted by $\mathcal{G}_+, \mathcal{G}_- \subseteq [n]$, and are of size $n_+ := |\mathcal{G}_+|, n_- := |\mathcal{G}_-|$, respectively. SPP takes as input a set of n labeled sequences:

$$\{(\boldsymbol{g}_i, y_i)\}_{i \in [n]},$$

where $\boldsymbol{g}_i$ represents the ith sequence/passage of play. Each sequence $\boldsymbol{g}_i$ takes a label from $y_i \in \{\pm 1\}$ and can be written as

$$\boldsymbol{g}_i = \langle g_{i1}, g_{i2}, \ldots, g_{iT(i)} \rangle, i \in [n],$$

where g_{it} is the tth symbol of the ith sequence, which takes one of the event symbols in $\mathcal{S}$, and $T(i)$ is the length of the ith sequence (i.e., $T(i)$ is the number of events in sequence $\boldsymbol{g}_i$). Patterns of play are denoted by $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots$, each of which is also a sequence of event symbols:

$$\boldsymbol{q}_j = \langle q_{j1}, q_{j2}, \ldots, q_{jL(j)} \rangle, j = 1, 2, \ldots,$$

where $L(j)$ is the length of pattern $\boldsymbol{q}_j$ for $j = 1, 2, \ldots$. The presence of subsequence $\boldsymbol{q}_j$ in sequence $\boldsymbol{g}_i$ is denoted by $\boldsymbol{q}_j \sqsubseteq \boldsymbol{g}_i$. The set of all possible patterns contained in any sequence $\{\boldsymbol{g}_i\}_{i\in[n]}$ is denoted as $\mathcal{Q} = \{\boldsymbol{q}_i\}_{i\in[d]}$, where d is the number of possible patterns ($\mathcal{Q}$ is very large in general, which is why the pruning and safe-screening mechanisms of SPP are useful).

Materials and Methods

Data

We obtained XML data generated from video that was tagged in Hudl Sportscode (`https://www.hudl.com/products/sportscode`) by the performance analyst of one of the teams in the Japan Top League competition (the team is not named for reasons of confidentiality). Written consent was obtained to use the data for research purposes. Seasons are comprised of a number of matches, matches are made up of sequences of play, which are, in turn, comprised of events. Our dataset consisted of all of this particular team's matches in their 2018 season for each of the opposition teams they faced. These matches consist of passages of play (i.e., sequences of events), however, rules need to be specified to decide the point at which these passages of play start and end. Initially, each match in the original dataset was one long sequence of events. One approach that we considered initially, which we used on other datasets, was to label match sequences based on whether the team won or lost the match. However, in our initial experiments, this did not produce interesting results since it is obvious that a greater number of scoring events will occur within match sequences labeled with wins, and so the discriminative patterns identified largely contained only contain these scoring events. Therefore, we generated a more granular dataset by specifying rules to delimit the match sequences into sequences representing individual passages of play (these rules are described in more detail in the following subsection). Each sequence was comprised of a series of events from 24 unique events (12 unique events for the team and opposition teams), based on the events the analyst had tagged in SportsCode. These events are listed in Table 2, and some are also depicted in Fig 2 (the XML data also contained a more granular level of data with a greater number of unique events, however, in order to reduce computational complexity, the higher level of granularity was considered).

Methods

Delimiting matches into sequences

First, each match sequence was delimited into passage of play event sequences (Fig 3). The rules to delimit matches into passages of play should ideally result in passages of play that begin and end at logical points in the match, e.g., when certain events occur, when play stops, or when possession changes (e.g., [43]), and should result in sequences which are neither overly long nor overly short. In this study, a passage of play was defined to start with either a kick restart, scrum, or line-out. These three events result in play temporarily stopping and therefore represent natural delimiters for our dataset. When a kick restart, scrum (except for a scrum reset where a scrum follows another

Table 2. Unique events in the original XML data. Events prefixed by "O-" are performed by the opposition team; those that are not are performed by the team.

event ID	event	event description
1	Restart Received	Team receives a kick restart made by the opposition team
2	Phase	Period between breakdowns (team in possession of the ball)
3	Breakdown	Team player is tackled, resulting in a ruck
4	Kick in Play	Kick within the field of play (rather than to touch) made by the team
5	Penalty Conceded	Team gives away a penalty, opposition may re-gain possession
6	Kick at Goal	Team attempts kick at goal
7	Quick Tap	Quick restart of play by the team following a free kick awarded to them
8	Line-out	Ball is thrown in by the team
9	Error	Mistake made by the team, e.g., lost possession, forward pass, etc.
10	Scrum	Set piece in which the forwards attempt to push the opposing team off the ball
11	Try Scored	Team places the ball down over opposition team's line (five points)
12	Line Breaks	Team breaches the opposition team's defensive line
13	O-Restart Received	Opposition team receives a kick restart made by the team
14	O-Phase	Period between breakdowns (opposition team in possession of the ball)
15	O-Breakdown	Opposition player is tackled, resulting in a ruck
16	O-Kick in Play	Kick within the field of play (rather than to touch) made by the opposition team
17	O-Penalty Conceded	Opposition team gives away a penalty, team may re-gain possession
18	O-Kick at Goal	Opposition team attempts kick at goal
19	O-Quick Tap	Quick restart of play by the opposition team following a free kick awarded to them
20	O-Line-out	Ball is thrown in by the opposition team
21	O-Error	Mistake made by the opposition team, e.g., lost possession, forward pass, etc.
22	O-Scrum	Set piece in which the forwards attempt to push the team off the ball
23	O-Try Scored	Opposition team places the ball down over the team's line (five points)
24	O-Line Breaks	Opposition team breaches the team's defensive line

scrum), or line-out occurs, this event becomes the first event in a new event sequence; otherwise, if a try is scored or a kick at goal occurs, a new passage of play also begins. Applying these rules (also shown in Fig 3) resulted in a delimited dataset consisting of 490 sequences, each made up of events listed in Table 2. At this stage, the delimited dataset was unlabelled, and the scoring events (try scored, kick at goal) for the team and opposition teams were contained in the sequences.

Experimental dataset creation and comparative approach

The delimited sequence data described above was then divided into two datasets. In the first, which we call the scoring dataset, we consider the case in which the sequences are from the team's scoring perspective. In this dataset, the label $y_i = +1$ represents points being scored or attempted by the team. Note that while a try scored was certain in terms of points being scored, a kick at goal (depicted in the top-left of Fig 2) is not always successful and therefore may not result in points being scored. In our data, only the kick at goal being attempted (event id 6) was available—not whether the goal was actually successful or not. However, since it is more important to be able to identify points-scoring opportunities than whether or not the kick was ultimately successful (which is determined by the accuracy of the goal kicker), we assumed that all kicks at goal resulted in points being scored. In the scoring dataset, the label $y_i = +1$ was assigned to sequence i if a try was scored or a kick at goal was made by the team in that particular sequence. If no try was scored and no kick at goal was made by the

Fig 2. Key events in rugby matches. The photographs used as the original images are listed in parentheses. All of them are licensed under the unsplash.com license (https://unsplash.com/license). Top left: Kick at goal (https://unsplash.com/photos/xJSPP3H8XTQ); Bottom left: Line-out (https://unsplash.com/photos/CTEvFbFpVC8); Center top: Kick restart/Kick-off (https://unsplash.com/photos/OMdge7F2FyA); Center bottom: Scrum (https://unsplash.com/photos/y5H3_7OobJw); Top Right: Line break (https://unsplash.com/photos/XAlKHW9ierw); Middle Right: Beginning of a phase (https://unsplash.com/photos/fqrzserMsX4); Bottom Right: Breakdown (https://unsplash.com/photos/WByu11skzSc)

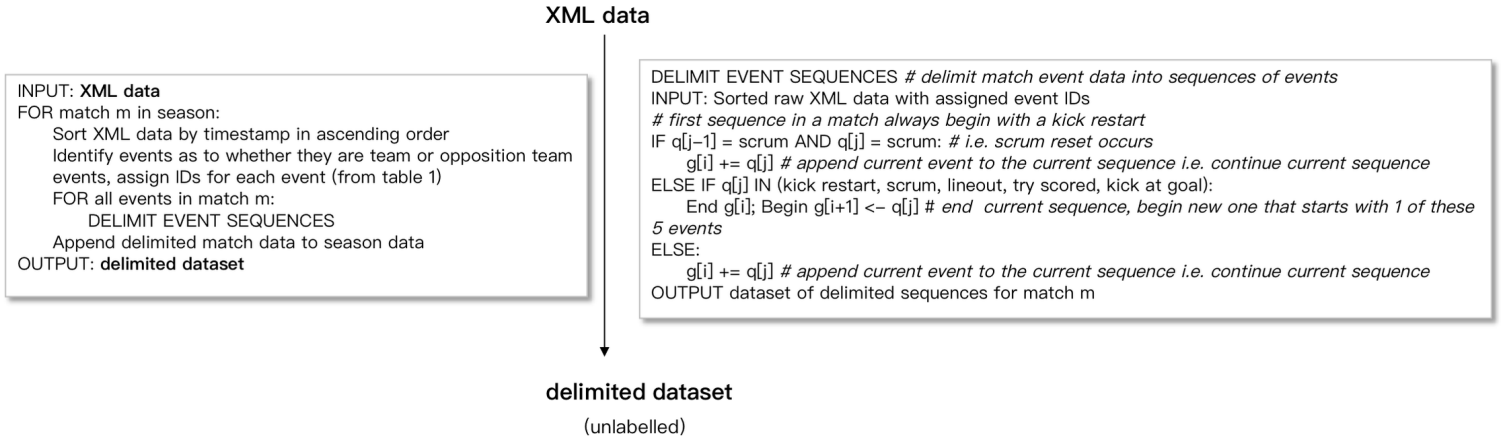


Fig 3. Illustration of the procedure to delimit the raw XML data into labelled sequences of events.

team in sequence i, the label $y_i = -1$ was assigned. Then, since the label now identified the scoring/not scoring outcome, the events that relate to the team scoring—Try scored (event ID = 11) and Kick at goal (event ID = 6)—were removed from the event sequences.

In the second, which we call the conceding dataset, we consider the case in which the sequences are from the team's conceding perspective, or equivalently, from the opposition teams' scoring perspective. In the conceding dataset, the label $y_i = +1$ was assigned to sequence i if a try was scored or a kick at goal was made by the *opposition* team in that particular sequence. If no try was scored and no kick at goal was made by the *opposition* team in sequence i, the label $y_i = -1$ was assigned. The list of events for the original delimited, scoring and conceding datasets are presented in Table 3. Then, since the label now identified the scoring/not scoring outcome, the events that relate to the *opposition* team scoring—Try scored (event ID = 11) and Kick at goal (event ID = 6)—were removed from the event sequences.

The process to create the scoring and conceding datasets from the original delimited dataset is shown in the upper half of Fig 4.

Table 3. Event lists for the original, scoring and conceding datasets.

event ID	original	scoring	conceding
1	Restart Received	Restart Received	Restart Received
2	Phase	Phase	Phase
3	Breakdown	Breakdown	Breakdown
4	Kick in Play	Kick in Play	Kick in Play
5	Penalty Conceded	Penalty Conceded	Penalty Conceded
6	Kick at Goal		Kick at Goal
7	Quick Tap	Quick Tap	Quick Tap
8	Line-out	Line-out	Line-out
9	Error	Error	Error
10	Scrum	Scrum	Scrum
11	Try Scored		Try Scored
12	Line Breaks	Line Breaks	Line Breaks
13	O-Restart Received	O-Restart Received	O-Restart Received
14	O-Phase	O-Phase	O-Phase
15	O-Breakdown	O-Breakdown	O-Breakdown
16	O-Kick in Play	O-Kick in Play	O-Kick in Play
17	O-Penalty Conceded	O-Penalty Conceded	O-Penalty Conceded
18	O-Kick at Goal	O-Kick at Goal	
19	O-Quick Tap	O-Quick Tap	O-Quick Tap
20	O-Line-out	O-Line-out	O-Line-out
21	O-Error	O-Error	O-Error
22	O-Scrum	O-Scrum	O-Scrum
23	O-Try Scored	O-Try Scored	
24	O-Line Breaks	O-Line Breaks	O-Line Breaks
label	-	Points Scored	O-Points Scored
	n=490	n_+=86, n_-=404	n_+=44, n_-=446

The SPP algorithm (software is available at `https://github.com/takeuchi-lab/SafePatternPruning`) was applied to the scoring and conceding datasets.

As a basis for comparison, we compared the patterns (q_js) obtained by SPP with

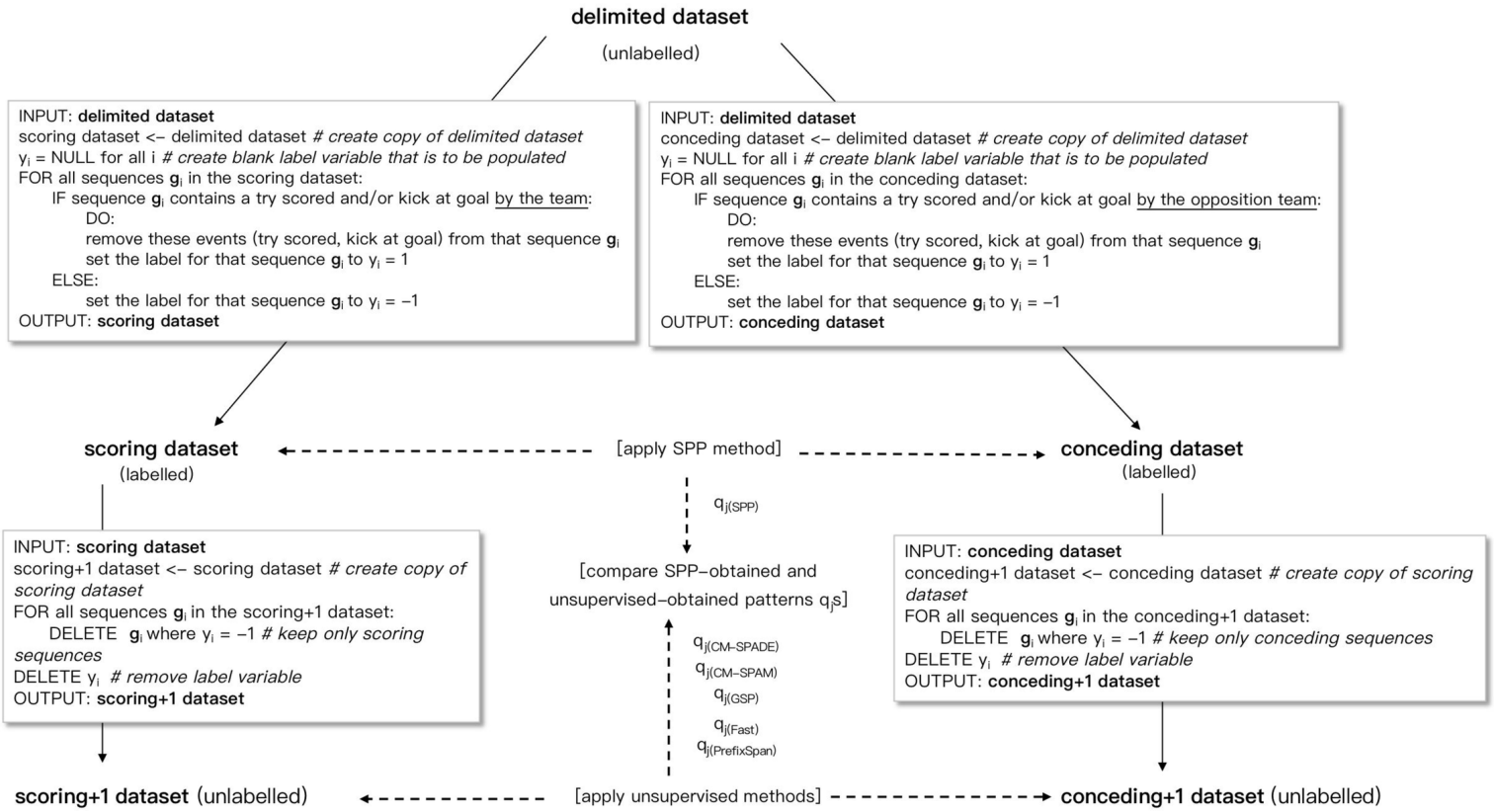


Fig 4. Illustration of dataset creation and experimental approach. Illustration of the procedures to create the datasets from the original delimited dataset to be used in the experiments and to compare the unsupervised and supervised SPM methods.

those obtained by the following unsupervised methods: PrefixSpan, CM-SPAM, CM-SPADE, GSP and Fast. The SPMF pattern mining package [44] (v2.42c) was used to apply the five unsupervised SPM methods to our dataset. Since the unsupervised methods use unlabeled data, although support values of the patterns of play can be obtained (how many times a particular pattern occurred in the dataset), weights for the patterns cannot. For a more fair comparison between the unsupervised methods and SPP, we assume that the unsupervised methods have prior knowledge of the sequence labels. Thus, the unsupervised methods were applied to subsets of the scoring and conceding datasets, partitioned based on the label. The first subset, which we call the "scoring+1" dataset, contained only the sequences in which the team actually scored, and the second, the "conceding+1" dataset, contained only the sequences in which the team actually conceded (i.e., the opposition team scored).

The dataset creation process and comparative approach is presented in Fig 4.

Obtaining pattern weights with safe pattern pruning

As mentioned, our data consisted of sequences comprised of events from Table 2, labeled with an outcome: either +1 or -1, e.g.

```
-1 22 22 17
+1 8 11 2 6
-1 1 2 3 2 9
```

```
-1 20 21
-1 10 10 2 3 2 3 2 3 2 3 2 17
+1 8 11 2 6
-1 1 2 3 2 3 9
-1 22 16 2
-1 13 14 16
...
```

We used SPP to identify patterns that discriminate between outcome +1 and outcome -1. For instance, in the dataset above, it would appear that subsequence [2 3 2] is potentially a discriminative pattern, since it appears in three sequences that are labeled with -1 but does not appear in any sequences that are labeled with 1. Pattern [11 2 6] also appears to be a discriminative pattern since it appears in two sequences with label 1 and in none labeled with -1. In SPP, after performing safe screening and pruning, each remaining pattern in a sequence is multiplied by weights, e.g., w_1[2 3 2] + w_2[11 2 6]..., and then an optimization model solves for these weights.

SPP uses a classifier based on a sparse linear combinations of patterns, which can be written as

$$f(\boldsymbol{g}_i; \mathcal{Q}) = \sum_{\boldsymbol{q}_j \in \mathcal{Q}} w_j I(\boldsymbol{q}_j \sqsubseteq \boldsymbol{g}_i) + b, \tag{1}$$

where $I(\cdot)$ is an indicator function that takes the value 1 if sequence $\boldsymbol{g}_i$ contains pattern $\boldsymbol{g}_i$ and 0 other otherwise; and $w_j \in \mathbb{R}$ and $b \in \mathbb{R}$ are parameters of the linear model that are estimated by solving the following minimization problem (as well as its dual maximization problem):

$$\min_{\boldsymbol{w},b} \sum_{i \in [n]} \ell(y_i, f(\boldsymbol{g}_i; \mathcal{Q})) + \lambda \|\boldsymbol{w}\|_1, \tag{2}$$

where $\boldsymbol{w} = [w_1, \ldots, w_d]^\top$ is a vector of weights, ℓ is a loss function and $\lambda > 0$ is a regularization parameter that can be tuned by cross-validation. Note that, due to the permutations in terms of the number of potential patterns of play, the size of $\mathcal{Q}$ is large in general. However, SPP's pruning criterion reduces the size of $\mathcal{Q}$ by removing unnecessary patterns from the original pattern tree. The minimization problem (1) was, in the present study, solved with an L1-regularised L2-Support Vector Machine (the default option in the S3P classifier command line options `https://github.com/takeuchi-lab/S3P-classifier`), with 10-times 10-fold cross-validation used to tune the regularization parameter, lambda. The maximum pattern length parameter (option -L in the S3P classifier command line options) was set to 20. The feature vector $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$ is defined for the ith sequence $\boldsymbol{g}_i$ as

$$x_{ij} = I(\boldsymbol{q}_j \sqsubseteq \boldsymbol{g}_i), \qquad j = 1, \ldots, |\mathcal{Q}|. \tag{3}$$

In other words, the feature vectors, $\boldsymbol{x}_i = [I(\boldsymbol{q}_1 \sqsubseteq \boldsymbol{g}_i), I(\boldsymbol{q}_2 \sqsubseteq \boldsymbol{g}_i), \ldots, I(\boldsymbol{q}_d \sqsubseteq \boldsymbol{g}_i)]$, are binary variables that take the respective values 1 or 0 based on whether or not pattern $\boldsymbol{q}_j$ is contained within sequence $\boldsymbol{g}_i$. In a two-class problem, the squared hinge-loss function $\ell(y, f(\boldsymbol{x}_i)) = \max\{0, 1 - yf(\boldsymbol{x}_i)\}^2$ is used, and the optimization problem (2) becomes:

$$\min_{\boldsymbol{w},b} \sum_{i \in [n]} \max\left\{0, 1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\right\}^2 + \lambda \|\boldsymbol{w}\|_1, \tag{4}$$

Discriminative patterns are those that have positive weights (in absolute terms) in the optimal solution to (4), i.e., those patterns that have weights that are non-zero. As

Table 4. Descriptive statistics for the scoring and conceding datasets

	scoring	conceding
Mean	10.6	10.8
Standard deviation	7.8	7.9
Minimum	2	2
25th percentile	5	5
Median	8	8
75th percentile	15	15
Maximum	48	48
Skewness	1.3	1.4

mentioned, SPP uses safe screening to remove weights that are guaranteed to be zero at the optimal solution, prior to solving the optimization problem (see S2 Appendix for more details.

In the results section below, in order to exclude patterns that may have occurred merely by chance, we have not reported obtained patterns (q_js) that had support values of less than five. In the case of the patterns obtained by the unsupervised methods, the top five patterns with the largest support values are reported. In the case of the SPP-obtained patterns, the top five patterns with the largest positive w_j values were recorded. In addition, we restricted our analysis to patterns of play that had the highest positive weights. For the scoring dataset, this means the patterns that had a positive contribution to the team scoring. For the conceding dataset, this means the patterns that had a positive contribution to opposition teams scoring. In other words, for the sake of brevity, we did not consider the patterns that had the highest contribution to "not scoring" and "not conceding."

Results and Discussion

Analysis of sequence lengths

The average sequence length in the scoring and conceding datasets was 10.6 and 10.8 events, respectively. The shortest sequence in both datasets contained two events, and the longest contained 48 events (Table 4). The slight difference in average sequence length between the scoring and conceding datasets is a result of the removal of the try scored and kick at goal events from the sequences in order to create the sequence outcome label (as was mentioned in the Materials and Methods section above). The sequence length distributions (Fig 5) are positively skewed and, based on Shapiro-Wilk tests, non-normal. By comparing these distributions, it is clear that the number of sequences in which points were scored was higher in the scoring dataset than the conceding dataset, which reflects this particular team's strength in the 2018 season. From the team's scoring perspective, 86 out of the 490 passages of play (18%) resulted in points being scored by the team, while from the team's conceding perspective, 44 out of the 490 passages of play (9%) resulted in points being conceded. The sequences in which the team scored were slightly longer, containing an average of 12.8 events compared to sequences in which the team didn't score, which contained an average of 10.2 events. The sequences in which the team conceded and did not concede contained an average of 11.2 events and 10.8 events, respectively.

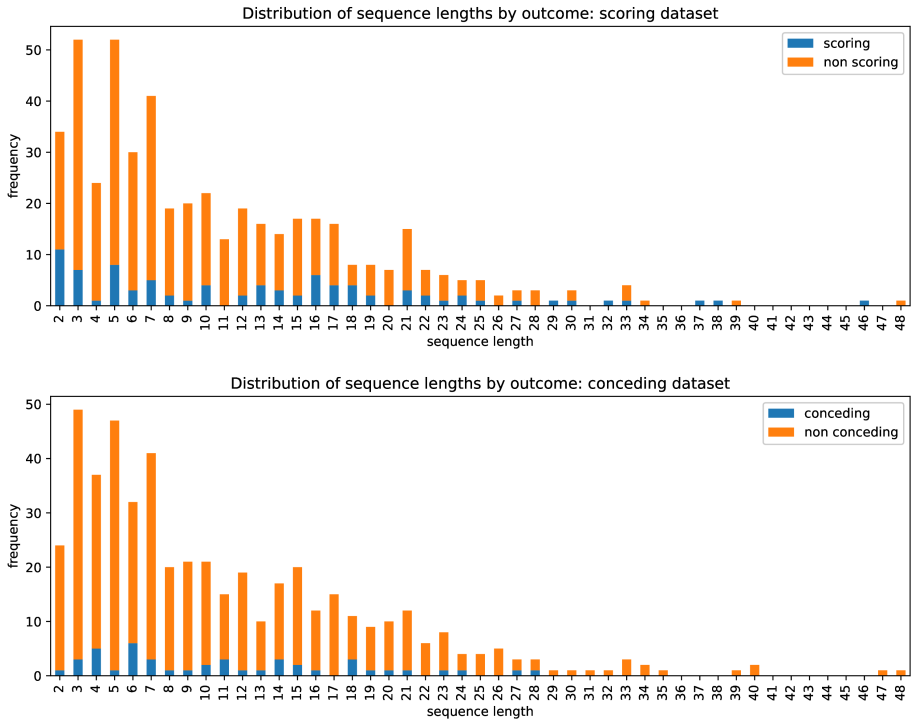


Fig 5. Sequence length distributions. Distribution of sequence lengths by points-scoring outcome for the scoring and conceding datasets. Sequence length is defined as the number of events in each sequence (excluding the outcome label).

Identification of important patterns of play using SPP

SPP initially obtained 93 patterns when applied to the scoring dataset, of which 75 had support of five or higher. Of these 75 patterns of play, 38 had a positive weight ($w_j > 0$). The 75 patterns with minimum support of five contained an average of 4.5 events, and the 38 patterns with positive weights contained an average of 5.4 events. The longest obtained pattern in the scoring dataset contained 16 events.

Applying SPP to the conceding dataset resulted in a total of 72 patterns, of which 51 had support of five or higher. Of these 51 patterns of play, 31 had a positive weight ($w_j > 0$). The 51 patterns with minimum support of 5 contained an average of 3.8 events, and the 31 patterns with positive weights contained an average of 4.4 events. The longest obtained pattern in the conceding dataset contained 15 events.

The five patterns that discriminated the most between scoring and non-scoring outcomes (i.e., the patterns with the largest positive weight values) were obtained by applying SPP to the scoring dataset, and are listed along with their weight values and odds ratios in Table 5. In the results tables, the notation $[p] \times n$ indicates that pattern p is repeated n times. We also include the odds ratio (OR) for each pattern (simply the exponential of the weight), which aids in interpretation by providing a value that compares the cases where a sequence contains a particular pattern, and when it does not.

The pattern in the scoring dataset with the highest weight value (0.919), which discriminated the most between scoring and non-scoring sequences, was a pattern consisting of a single line break event (event id 12). The OR for the line break pattern is $e^{0.919} = 2.506$, meaning that the team is 2.5 times more likely to score when a line break is made in a sequence of play than if a line break is not made in a sequence of

play. This makes sense since line breaks, which involve breaking through an opposition team's line of defense (see the top-right image in Fig 2), generally advance the attacking team forward and are thus expected to create possible scoring opportunities. A line-out followed by phase play (8 2) was the second most discriminative pattern between scoring and not scoring, with a weight of 0.808 and an OR of 2.242, indicating that the team is 2.2 times more likely to score when a line-out followed by a phase occurs in a sequence of play than if it does not. The third most discriminative pattern, 2 3 4 2 3 ($w = 0.796$, $OR = 2.217$), can be interpreted as a kick in play being made by the team and being re-gathered by the team, thus resulting in retained possession. The OR indicates that the team is 2.2 times more likely to score when this pattern occurs in a sequence of play than if it does not. The fourth most discriminative pattern, 2 3 2 3 2 3 2 3 4 ($w = 0.732$, $OR = 2.079$), represents four repeated phase-breakdown plays by the team, followed by the team making a kick in play. This pattern thus involves repeated retaining of possession before the team presumably gaining territory in the form of a kick. The OR indicates that the team is 2.1 times more likely to score when this pattern occurs in a sequence of play than if it does not. The fifth most discriminative pattern, 13 14 15 14 15 16 14 2 3 ($w = 0.710$, $OR = 2.033$), can be interpreted as the opposition team receiving a kick restart made by the team, attempting to exit their own territory via a kick but not finding touch, thus giving the ball back to the team from which they can potentially build phases and launch an attack. The OR indicates that the team is twice as likely to score when this pattern occurs in a sequence of play than if it does not.

Table 5. Top five SPP-obtained patterns that discriminated the most between scoring and non-scoring outcomes.

pattern (q_j)	pattern description	support	weight	OR
12	Line break	77	0.919	2.506
8 2	Line-out, Phase	71	0.808	2.242
2 3 4 2 3	Phase, Breakdown, Kick in play, Phase, Breakdown	9	0.796	2.217
2 3 2 3 2 3 2 3 4	[Phase, Breakdown]×4, Kick in play	9	0.732	2.079
13 14 15 14 15 16 14 2 3	O-Restart received, [O-Phase, O-Breakdown]×2, O-Kick in play, Phase, Breakdown	6	0.710	2.033

The five patterns that discriminated the most between conceding and non-conceding outcomes were obtained by applying SPP to the conceding dataset, and are listed along with their weights and ORs in Table 6. A line break (event ID 24) ($w = 0.613$, $OR = 1.846$) being made by the opposition team was the pattern that discriminated most between sequences in which the team conceded (i.e., the opposition team scored) and sequences in which the team did not concede (i.e., the opposition team did not score). In other words, a line break by the opposition team was the pattern that discriminated the most between the group of sequences in which the opposition team scored and the group of sequences in which the opposition team did not score. The OR of 1.8 indicates that the opposition team is 1.8 times more likely to score when they make a line break in a sequence of play than if they do not. The weight (and OR) for the line break pattern was not as large as in the scoring dataset ($w = 0.919$ vs. $w = 0.613$), which suggests that the team had strong defense in this particular season. The second most discriminative pattern between conceding and non-conceding outcomes, 14 9 15 ($w = 0.392$, $OR = 1.479$), can be interpreted as the opposition team being in possession of the ball, the team making some form of error, and the opposition team regaining possession. The opposition team is 1.5 times more likely to score when this pattern occurs in a sequence of play than if it does not. The third most discriminative pattern between conceding and non-conceding outcomes was an opposition team line-out ($w = 0.357$, $OR = 1.428$). The opposition team is 1.4 times

more likely to score if they have a line-out in a sequence of play than if they do not. The fourth ($w = 0.339$, $OR = 1.403$) and fifth ($w = 0.261$, $OR = 1.299$) most discriminative patterns for the conceding dataset represent repeated phase and breakdown play, which results in retained possession and the building of pressure. The fifth pattern, for example, is one in which the opposition team makes over six repeated consecutive phases and breakdowns.

Table 6. Top five SPP-obtained patterns that discriminated the most between conceding and non-conceding outcomes.

event id pattern (q_j)	pattern description	support	weight	OR
24	O-Line break	32	0.613	1.846
14 9 15	O-Phase, Error, O-Breakdown	10	0.392	1.479
20	O-Line-out	86	0.357	1.428
15 15 14 15	O-Breakdown, O-Breakdown, O-Phase, O-Breakdown	5	0.339	1.403
15 14 15 14 15 14 15 14 15 14 15 14 15	[O-Breakdown, O-Phase]×6, O-Breakdown	16	0.261	1.299

Comparing the SPP-obtained patterns with those obtained by the unsupervised methods

The five patterns with the highest support for the scoring+1 and conceding+1 datasets, obtained by applying each of the five unsupervised methods (PrefixSpan, CM-SPAM, CM-SPADE, GSP and Fast), are shown in Tables 7 and 8 , respectively.

Table 7. Top five PrefixSpan-obtained patterns with the largest support: scoring+1 dataset.

PrefixSpan	CM-SPAM	CM-SPADE	GSP	Fast	support
2	2	2	2	2	84
2 3	3	3	3	3	60
3	2 3	2 3	2 3	2 3	60
2 2	2 2	3 2	2 2	2 2	59
2 3 2	2 3 2	2 2	3 2	3 2	59

Table 8. Top five PrefixSpan-obtained patterns with the largest support: conceding+1 dataset.

PrefixSpan	CM-SPAM	CM-SPADE	GSP	Fast	support
14	14	14	14	14	39
14 15	15	15	15	15	33
15	14 15	14 15	14 15	14 15	33
14 14	14 14	15 14	14 14	14 14	29
14 15 14	14 15 14	14 14	15 14	15 14	29

Tables 7 and 8 show that only common events and patterns were detected by the unsupervised methods, i.e., patterns containing breakdowns, phases, or both. Repeated breakdown and phase play means that a team can generally retain possession of the ball and build pressure pressure (see the middle and bottom images on the right-hand side of Fig 2). While some of the patterns identified by SPP also contained repeated breakdown and phase play, they were generally longer and also contained other events.

The patterns obtained by the unsupervised methods are not particularly useful for coaches or performance analysts since they merely reflect common, repeated patterns rather than interesting patterns. By using passage of play event sequences that are labelled with scoring or conceding outcomes, through the computed weights, SPP is also able to provide a measure of the degree to which particular patterns discriminate between these outcomes. In addition, compared to the unsupervised methods, the supervised SPP method obtained a greater variety of patterns of play (i.e., not only those containing breakdowns and/or phases) and also discovered more sophisticated patterns that can be readily interpreted by coaches or performance analysts.

Discussion

By considering both the scoring and conceding perspectives of the team, insight was able to be obtained that would be useful to both the team as well as opposition teams that are due to play the team. For both the team and their opposition teams during the 2018 season, line breaks were found to be most associated with scoring. For both the team and their opposition teams, line-outs were found to be more beneficial in generating scoring opportunities than scrums. This result is consistent with [41], who found that line-outs followed by a driving maul are common approaches to scoring tries (albeit in a different competition, Super Rugby), and with [45], who found that around one-third of tries came from line-outs in the Japan Top League in 2003 to 2005—the highest of any try source. As well as creating line-outs or perhaps prioritising them over scrums, for opposition teams playing the team, effective strategies may include aiming to maintain possession through repeated phase-breakdown play (over six repetitions), shutting down the team's ability to regain kicks, and ensuring that touch is found on exit plays from kick restarts made by the team.

As mentioned, compared to the unsupervised methods, the supervised SPP method obtained a greater variety of patterns that consisted of a greater variety of events. The unsupervised methods only generated patterns consisting of either phases, breakdowns, or both, which are very frequent and repetitive patterns but are not of use to coaches or performance analysts. As well as containing a greater variety of events, the patterns obtained by SPP were more complex in terms of the patterns of play that were identified. For instance, through one of the SPP-obtained patterns, an opposition team could identify that they could be punished by the team for a failed exit play. The pattern involving the team making and regaining a kick in play (which was shown to discriminate between scoring and non-scoring outcomes for the team) is another example of a complex pattern of play that was identified by SPP. The superiority of SPP over the unsupervised methods is likely due to the discriminative nature of SPP as well as the safe screening and pattern pruning mechanisms of SPP, which prune out irrelevant sequential patterns and model weights in advance.

Conclusions and Future work

In this study, a supervised sequential pattern mining (SPM) method called safe pattern pruning (SPP) was applied to data from professional rugby union in Japan that consisted of sequences in the form of passages of play that were labelled with points scoring outcomes. The obtained results suggest that the SPP model was useful in detecting complex patterns (patterns of play) that are important to scoring outcomes. SPP was able to identify relatively sophisticated, discriminative patterns of play, which made sense when interpreted, and which are potentially useful for coaches and performance analysts for own- and opposition-team analysis in order to identify vulnerabilities and tactical opportunities. The approach highlighted the potential utility

of supervised SPM as an analytical framework for performance analysis in sport, and more specifically, the potential usefulness of SPM methods for performance analysis in rugby.

Although the results obtained are encouraging, a limited amount of data from one sport was used. Also, spatial information such as field position was not available in the data, and this may have improved the analysis. Although the team that performed a particular event was used in our analysis, which player performed particular events was not considered—this may be interesting to investigate in future work. One limitation of SPP is that, although it considers the order of events within the sequences, the method does not consider the order of sequences within matches, which could also be informative (e.g., a particular pattern occurring in the second half of a match may be more important than if it occurs in the first half). Furthermore, although SPP was useful for the specific dataset in this study, its usefulness is to some degree dependent on the structure of the input data and the specific definition of the sequences and labels. For instance, as mentioned earlier, applying the approach to a dataset that consists of entire matches as sequences and match win/loss outcomes does not tend to produce interesting results since it is self-evident that sequences that contain more scoring events will be more associated with wins, and so SPP would simply pick up the scoring events in such datasets. In future work, it would be interesting to apply the approach to a larger amount of data from rugby, as well as to similarly structured datasets in other sports in order to confirm its efficacy.

S1 Dataset. The delimited sequence data that is described in this paper is available on GitHub: `https://github.com/rorybunker/rugby-sequences`

S2 Appendix. Safe Screening and Regularization Path Initialization. Some weights are removed prior to solving (4) using safe screening, which corresponds to finding j such that $w_j = 0$ in the optimal solution $\boldsymbol{w}^* := [w_1^*, \ldots, w_d^*]^\top$ in the optimization problem (4). Such w_j do not affect the optimal solution even if they are removed prior to solving the optimization problem. In the optimal solution, the $\boldsymbol{w}^*$ of the optimization problem (2), a set of j such that $|w_j^*| > 0$ is called the active set, and is denoted as $\mathcal{A} \subseteq \mathcal{Q}$. In this case, even if only the patterns included in $\mathcal{A}$ are used, the same optimal solution—as when using all patterns—can be obtained. Thus, if one solves

$$(\boldsymbol{w}'^*_{\mathcal{A}}, b'^*) := \operatorname*{argmin}_{\boldsymbol{w}, b} \sum_{i \in [n]} \ell(y_i, f(\boldsymbol{g}_i; \{\boldsymbol{q}\}_{i \in \mathcal{A}})) + \lambda \|\boldsymbol{w}\|_1, \tag{5}$$

then it is guaranteed that $\boldsymbol{w}^* = \boldsymbol{w}'^*_{\mathcal{A}}$ and $b^* = b'^*$.

In practice, the λ parameter is found based on a model selection technique such as cross-validation. In model selection, a sequence of solutions, a so-called regularization path, with various penalty parameters must be trained. The regularization path of the problem (2), $\{\lambda_0, \lambda_1, \ldots, \lambda_K\}$, is usually computed with decreasing λ, because sparser solutions are obtained for larger λ.

To compute the regularization path, the initial values are set to $\boldsymbol{w}^* \leftarrow \boldsymbol{0}$, $b^* \leftarrow \bar{y}$ (where $\bar{y}$ is the sample mean of $\{y_i\}_{i \in [n]}$) and $\lambda_0 \leftarrow \lambda_{\max}$ (see [34] for how $\lambda_{\max}$ is calculated, and for further details of the SPP method as well as its safe-screening mechanism and pruning criterion).

References

1. Hughes MD, Bartlett RM. The use of performance indicators in performance analysis. Journal of sports sciences. 2002 Jan 1;20(10):739-54.

2. Agrawal R, Srikant R. Mining sequential patterns. In: Proceedings of the eleventh international conference on data engineering. 1995 Mar 6 (pp. 3-14). IEEE.

3. Mabroukeh NR, Ezeife CI. A taxonomy of sequential pattern mining algorithms. ACM Computing Surveys (CSUR). 2010 Dec 3;43(1): 1-41.

4. Wang K, Xu Y, Yu JX. Scalable sequential pattern mining for biological sequences Proceedings of the thirteenth ACM international conference on Information and knowledge management. 2004 Nov 13 (pp. 178-187).

5. Ho J, Lukov L, Chawla S. Sequential pattern mining with constraints on large protein databases. In Proceedings of the 12th international conference on management of data (COMAD). 2005 (pp. 89-100).

6. Exarchos, TP, Papaloukas, C, Lampros, C, Fotiadis, DI. Mining sequential patterns for protein fold recognition Journal of Biomedical Informatics. 2008 Feb 1;41(1):165-79

7. Hsu CM, Chen CY, Liu BJ, Huang CC, Laio MH, Lin CC, Wu TL. Identification of hot regions in protein-protein interactions by sequential pattern mining BMC bioinformatics. 2007 May;8(5):1-5.

8. Garboni C, Masseglia F, Trousse B. Sequential pattern mining for structure-based XML document classification. In International Workshop of the Initiative for the Evaluation of XML Retrieval. 2005 Nov 28 (pp. 458-468).

9. Feng J, Xie F, Hu X, Li P, Cao J, Wu X. Keyword extraction based on sequential pattern mining. In proceedings of the third international conference on internet multimedia computing and service. 2011 Aug 5 (pp. 34-38).

10. Xie F, Wu X, Zhu X. Document-specific keyphrase extraction using sequential patterns with wildcards. In: 2014 IEEE International Conference on Data Mining 2014 Dec 14 (pp. 1055-1060). IEEE.

11. Xie F, Wu X, Zhu X. Efficient sequential pattern mining with wildcards for keyphrase extraction. Knowledge-Based Systems. 2017 Jan 1;115:27-39.

12. Yap GE, Li XL, Philip SY. Effective next-items recommendation via personalized sequential pattern mining. In: International conference on database systems for advanced applications 2012 Apr 15 (pp. 48-64). Springer, Berlin, Heidelberg.

13. Salehi M, Kamalabadi IN, Ghoushchi MB. Personalized recommendation of learning material using sequential pattern mining and attribute based collaborative filtering. Education and Information Technologies. 2014 Dec;19(4):713-35.

14. Ceci M, Lanotte PF, Fumarola F, Cavallo DP, Malerba D. Completion time and next activity prediction of processes using sequential pattern mining. In International Conference on Discovery Science 2014 Oct 8 (pp. 49-61). Springer, Cham.

15. Wright AP, Wright AT, McCoy AB, Sittig DF. The use of sequential pattern mining to predict next prescribed medications. Journal of biomedical informatics. 2015 Feb 1;53:73-80.

16. Tsai CY, Lai BH. A location-item-time sequential pattern mining algorithm for route recommendation. Knowledge-Based Systems. 2015 Jan 1;73:97-110.

17. Tarus JK, Niu Z, Kalui D. A hybrid recommender system for e-learning based on context awareness and sequential pattern mining. Soft Computing. 2018 Apr;22(8):2449-61.

18. Fournier-Viger P, Lin JCW, Kiran RU, Koh YS, Thomas R. A survey of sequential pattern mining. Data Science and Pattern Recognition. 2017;1(1):54–77.

19. Yao H, Hamilton HJ, Butz, CJ. A foundational approach to mining itemset utilities from databases. In: Proceedings of the 2004 SIAM International Conference on Data Mining 2004 Apr 22 (pp. 482-486). Society for Industrial and Applied Mathematics.

20. Gan W, Lin JC, Zhang J, Fournier-Viger P, Chao H, Yu P. Fast utility mining on sequence data. IEEE transactions on cybernetics. 2020 Feb 28;51(2):487-500.

21. Yin J, Zheng Z, Cao L. USpan: an efficient algorithm for mining high utility sequential patterns. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2012 Aug 12 (pp. 660-668).

22. Wang J, Huang J, Chen Y. On efficiently mining high utility sequential patterns. Knowledge and Information Systems. 2016 Nov;49(2):597-627.

23. Lin JC, Srivastava G, Li Y, Hong T, Wang S. Mining High-Utility Sequential Patterns in Uncertain Databases. In: 2020 IEEE International Conference on Big Data (Big Data) 2020 Dec 10 (pp. 5373-5380). IEEE.

24. Srivastava G, Lin JC, Zhang X, Li Y. Large-scale high-utility sequential pattern analytics in Internet of things. IEEE Internet of Things Journal. 2020 Sep 25.

25. Wu JM, Srivastava G, Wei M, Yun U, Lin JC. Fuzzy high-utility pattern mining in parallel and distributed Hadoop framework. Information Sciences. 2021 Apr 1;553:31-48.

26. Srikant R, Agrawal R. Mining sequential patterns: Generalizations and performance improvements. In: International Conference on Extending Database Technology 1996 Mar 25 (pp. 1-17). Springer, Berlin, Heidelberg

27. Agarwal R, Srikant R. Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB 1994 Sep 12 (Vol. 1215, pp. 487-499).

28. Zaki, MJ. SPADE: An efficient algorithm for mining frequent sequences Machine learning. 2001 Jan;42(1):31-60.

29. Ayres J, Flannick J, Gehrke J, Yiu T. Sequential pattern mining using a bitmap representation. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2002 Jul 23 (pp. 429-435).

30. Pei J, Han J, Mortazavi-Asl B, Wang J, Pinto H, Chen Q, Dayal U, Hsu MC. Mining sequential patterns by pattern-growth: The prefixspan approach. IEEE Transactions on knowledge and data engineering. 2004 Oct 4;16(11):1424-40.

31. Fournier-Viger P, Gomariz A, Campos M, Thomas R. Fast vertical mining of sequential patterns using co-occurrence information. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining 2014 May 13 (pp. 40-52). Springer, Cham.

32. Salvemini E, Fumarola F, Malerba D, Han J. Fast sequence mining based on sparse id-lists. In International Symposium on Methodologies for Intelligent Systems 2011 Jun 28 (pp. 316-325). Springer, Berlin, Heidelberg.

33. Nakagawa K, Suzumura S, Karasuyama M, Tsuda K, Takeuchi I. Safe pattern pruning: An efficient approach for predictive pattern mining. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 1785-1794).

34. Sakuma T, Nishi K, Kishimoto K, Nakagawa K, Karasuyama M, Umezu Y, Kajioka S, Yamazaki SJ, Kimura KD, Matsumoto S, Yoda K. Efficient learning algorithm for sparse subsequence pattern-based classification and applications to comparative animal trajectory data analysis. Advanced Robotics. 2019 Feb 16;33(3-4):134-52.

35. Ghaoui LE, Viallon V, Rabbani T. Safe feature elimination for the lasso and sparse supervised learning problems. arXiv preprint arXiv:1009.4219. 2010 Sep 21.

36. La Puma I, de Castro Giorno FA. Ontology-Based Data Mining Approach for Judo Technical Tactical Analysis. In: The third international conference on computing technology and information management (ICCTIM2017) 2017 Dec 8 (p. 90).

37. Hrovat G, Fister Jr I, Yermak K, Stiglic G, Fister I. Interestingness measure for mining sequential patterns in sports. Journal of Intelligent & Fuzzy Systems. 2015 Jan 1;29(5):1981-94.

38. Decroos T, Van Haaren J, Davis J. Automatic discovery of tactics in spatio-temporal soccer match data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2018 Jul 19 (pp. 223-232).

39. Carter A, Potter G. The 1995 rugby world cup finals. 187 tries. In Hughes, M. (Ed.) (2001). Notational Analysis of Sport III. University of Wales Institute Cardiff. Cardiff, Wales UWIC 2001 (pp. 224-229).

40. van Rooyen KM, Noakes DT. Movement time as a predictor of success in the 2003 Rugby World Cup Tournament. International Journal of Performance Analysis in Sport. 2006 Jun 1;6(1):30-9.

41. Coughlan M, Mountifield C, Sharpe S, Mara JK. How they scored the tries: applying cluster analysis to identify playing patterns that lead to tries in super rugby. International Journal of Performance Analysis in Sport. 2019 May 4;19(3):435-51.

42. Watson N, Hendricks S, Stewart T, Durbach I. Integrating machine learning and decision support in tactical decision-making in rugby union. Journal of the Operational Research Society. 2020 Jul 31:1-2.

43. Liu T, Fournier-Viger P, Hohmann A. Using diagnostic analysis to discover offensive patterns in a football game. In: Recent Developments in Data Science and Business Analytics 2018 (pp. 381-386). Springer, Cham.

44. Fournier-Viger P, Gomariz A, Gueniche T, Soltani A, Wu CW, Tseng VS. Spmf: a java open-source pattern mining library. J. Mach. Learn. Res.. 2014 Jan;15(1):3389-93.

45. Sasaki K, Furukawa T, Murakami J, Shimozono H, Nagamatsu M, Miyao M, Yamamoto T, Watanabe I, Yasugahira H, Saito T, Ueno Y. Scoring profiles and defense performance analysis in Rugby Union. International Journal of Performance Analysis in Sport. 2007 Oct 1;7(3):46-53.

CHAPTER 6: PUBLICATION 3 - "PERFORMANCE INDICATORS CONTRIBUTING TO SUCCESS AT THE GROUP AND PLAY-OFF STAGES OF THE 2019 RUGBY WORLD CUP"

This study considered performance at the tournament stage level, particularly outcomes at the group and play-off tournament stages. The outcomes at these two stages of the tournament reflect the results of individual matches at these respective stages.

Various performance indicators were constructed using publicly available data from the official 2019 Rugby World Cup (RWC) tournament website. Performance indicators that led to success at the group and play-off stages of the tournament were identified using Wilcoxon's signed rank test and Cohen's d effect sizes, statistical techniques commonly used in similar prior studies' analyses. In addition, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) (Cohen, 1995), a rules-based machine learning algorithm that is fast to run and produces readily interpretable decision rules, was also applied. The decision rules generated by the RIPPER automatically identify relevant features for a tournament stage. Also, they provided a measure of feature importance by computing the percentage of winning and losing matches to which a specific decision rule applies. Another advantage of RIPPER is that it could identify cases where more than one performance indicator jointly contributed to a specific outcome, unlike the statistical techniques, which are applied in a univariate manner to each feature individually.

The results from Wilcoxon's signed rank test and Cohen's d effect sizes found that ball carry effectiveness, the percentage of ball carries that penetrated the opposition gain-line, and total metres gained (kick metres plus carry metres) contributed to success at both the group and play-off stages of the tournament. Notably, performance indicators that contributed to success during the group stages — dominating possession, making more ball carries and passes, winning more rucks, and making fewer tackles — did not contribute to success at the play-off stage. The results obtained using RIPPER suggested low ball carries and a low lineout success percentage jointly contributed to losing at the group stage, and having a greater number of carries penetrating the opposition team's gainline contributed to success. Surprisingly, having a fewer number of rucks was found to contribute to winning at the play-off stage of the tournament. Overall, the findings suggested the benefit of teams adapting their strategies when proceeding from the group stage to the play-off stage at the RWC. Compared with the calculation of Wilcoxon signed rank tests and Cohen's d effect sizes for

each of the performance indicator variables individually, RIPPER decision rules were quick to generate, easy to interpret, and did not require distributional assumptions of normality.

The significance of this work in terms of its contribution to the literature was in applying a technique for identifying key patterns from performance indicators. RIPPER was found to be helpful as an alternative to the commonly applied traditional statistical techniques in the field, the Wilcoxon's signed rank test and Cohen's d effect sizes, to identify key performance indicators (as well as combinations of performance indicators and their values) that discriminate between successful and unsuccessful outcomes. The suggested avenues for future work included incorporating contextual information and testing other interpretable machine learning methods.

Bunker, R.P., & Spencer, K. (2022). Performance indicators contributing to success at the group and play-off stages of the 2019 Rugby World Cup. Journal of Human Sport & Exercise, 17(3): 683-698. https://doi.org/10.14198/jhse.2022.173.18

Performance Indicators Contributing To Success At The Group And Play-Off Stages Of The 2019 Rugby World Cup

R. P. Bunker[a*] and K. Spencer[b*]

[a]RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, Japan 103-0027. ORCID: 0000-0001-9243-7063

[b]Sport Performance Research Institute New Zealand, Auckland University of Technology, Auckland 0632, New Zealand. ORCHID: 0000-0002-3188-2179

Abstract

Performance indicators that contributed to success at the group and play-off stages of the 2019 Rugby World Cup were analysed using publicly available data obtained from the official tournament website using both a non-parametric statistical technique, Wilcoxon's signed rank test, and a decision rules technique from machine learning called RIPPER. Our statistical results found that ball carry effectiveness (percentage of ball carries that penetrated the opposition gain-line) and total metres gained (kick metres plus carry metres) were found to contribute to success at both stages of the tournament and that indicators that contributed to success during group stages (dominating possession, making more ball carries, making more passes, winning more rucks, and making less tackles) did not contribute to success at the play-off stage. Our results using RIPPER found that low ball carries and a low lineout success percentage jointly contributed to losing at the group stage, while winning a low number of rucks and carrying over the gain-line a sufficient number of times contributed to winning at the play-off stage of the tournament. The results emphasise the need for teams to adapt their playing strategies from group stage to the play-off stage at the tournament in order to be successful.

Keywords

Decision Rules; Rugby Union; Sport Performance Analysis; Machine Learning; Non-parametric Statistics; RIPPER

1. Introduction

The Rugby World Cup (RWC) is a major global sports event that is held every four years and involves the top 20 rugby countries. The RWC was first held in 1987 yet the 2019 tournament was first one to be held in Japan. South Africa captured the title for a third time, equalling New Zealand's record. Matches during the group stage of the 2019 RWC were noticeably closer than at previous tournaments, perhaps reflective of the narrowing performance gap between Tier one and Tier two (higher ranked and lower ranked) nations.

Over the past few decades, rugby has evolved considerably, with World Rugby implementing various law changes with the objective of promoting changes in playing behaviour such as increasing running with the ball in order to increase the attractiveness of the game to spectators. For instance, compared to the 1995 RWC, ball-in-play time at the 2011 RWC increased by 33%, the number of passes made increased by ~50%, the frequency of rucks/mauls more than doubled, the number of kicks was reduced by 50%, and the frequency of scrums per match decreased from 27 to 17 (Vaz, Vasilica, Kraak & Arrones, 2015). This suggests that the rule changes have impacted on performance.

Performance indicators (M.D. Hughes & Bartlett, 2002) have been constructed using data derived from notational analysis and analysed for some time in rugby union, primarily using descriptive statistics as well as formal statistical methods. Early studies on RWC tournaments made use of notational analysis on match videos, were largely descriptive and generally did not make use of formal statistical methods. M. Hughes and White (1997) investigated differences in the patterns of play of the forwards of successful and unsuccessful teams at the 1991 RWC. They found that the forwards of successful teams dominated lineouts by having a greater variety of lineout options, were technically superior in winning more scrums per match, and were dominant at ruck and maul time. Stanhope and M. Hughes (1997) examined team performances from the 1991 RWC, investigating the different ways points were scored in terms of their tactical significance to successful teams. Although similar patterns of play were generally observed between successful and unsuccessful teams, the noticeable differences were that the successful teams were superior at ruck time and in their kicking performance; in particular, successful teams kicked into danger areas of the field from which to launch attacks. The authors also found that dominant rucking and kicking games also produced more penalties for successful teams, which they were able to take advantage of in danger areas of the field. McCorry, Saunders, O'Donoghue, and Murphy (2001) found that the possession gain to loss, that is, turnovers won or lost, reflected the final ranking of the semi-finalists at the 1995 RWC. Hunter and O'Donoghue (2001) assessed positive and negative aspects of attacking and defensive play, changes in possession, and methods used to gain territory at the 1999 RWC. They found that winning and losing sides differed in the frequency with which they penetrated the opposition's last third of the field, and the frequency of attacking plays in which they outflanked the opposition team. In another study using coded match footage from the 48 tournament matches, it was found that ruck frequency was a strong predictor of success at the 2007 RWC (van Rooyen, Diedrick, & Noakes, 2010). This study was largely descriptive and didn't make use of formal statistical methods. Play-off and group stage matches were considered separately, a similar approach to the present study. It was found that 100% of the play-off matches were won by teams with a lower number of rucks; however, higher ruck frequency was associated with success during the group stage matches. This suggested that the tournament format may influence a team's tactics, and the PIs that are important to success may differ at different stages of the tournament. Vaz et al. (2015) used video from the 12 matches involving New Zealand and France at the 2011 RWC, along with data obtained from the official RWC website and rugbystats.com.au. The data was obtained from coded match footage. They found that there were significant differences in PIs for these two teams. New Zealand's points resulted mostly from tries,

while France scored their points mostly from penalties, and there were differences in the performance levels between match halves.

Over the past two decades, data from RWC tournaments has increasingly been made available online, and some studies have augmented data from notational analysis with online-sourced data. There has also been increased focus on ensuring the reliability and validity of data obtained via notational analysis, and on making use of formal statistical methods. In a study by van Rooyen, Lambert, & Noakes (2006), PI data was obtained from the International Rugby Board (IRB) official website and these were augmented with other variables related to movements, which were derived from video analysis of match play. The aim of the study was to explain the performance of four teams at the 2003 RWC. The authors compared South Africa, who lost at the quarter final stage, with the top three teams (England, New Zealand and Australia). The dataset that compared the performance of the teams was analysed using the Kruskal-Wallis test. The data related to the field location of movements were analysed for statistical significance using the Chi-squared test, a one-tailed t-test, and Spearman's rank order correlation coefficients. The most important variables found were the amount of time teams were in possession of the ball, the number of points scored in the second half, and the loss of possession in areas of the field in which the opposition team was likely to score from. In a study using video analysis and Chi-squared tests, van Rooyen and Noakes (2006) found that movement time was an important predictor of success at the 2003 RWC. In particular, they found that a team's ability to construct movements that lasted longer than 1 minute and 20 seconds was important in determining where teams finished at the tournament. Bishop and Barnes (2013) investigated PIs that discriminated between winning and losing teams at the play-off stage of the 2011 RWC. Their data was obtained via coded match footage, which was then checked for intra-observer by a second analyst coding the same matches, as well as re-analysis reliability two weeks later. Several of the PIs were non-normally distributed, thus, the non-parametric Wilcoxon signed rank test (Wilcoxon, 1945) was used. Statistically significant differences between two performance indicators were found, with winning teams conceding a higher percentage of their penalties between the 50 metre and the opposition 22 metre line, and winning teams kicking the ball out of hand more than losing teams. Notably, they also found that successful teams played a more territory-based game as opposed to a possession-based approach. A. Hughes, Barnes, Churchill, and Stone (2017) used data derived from coded match footage, which was coded by two analysts for reliability, and compared PIs that discriminated between winning and losing in the play-off stages of the men's 2015 RWC and 2014 women's RWC. The Shapiro-Wilkes test was used to check normality, and it was found that 91% of the variables were normally distributed; therefore, the parametric two-way mixed ANOVA was used to test for differences between winning and losing teams as well as between genders. They found that in the men's 2015 RWC, winning teams kicked a higher percentage of possession in the opposition 22 to 50 metre zone of the field, for the purpose of creating territory-related pressure (women's was found to be more possession oriented), and the percentage of lineouts won on the opposition throw was found to discriminate between winners and losers.

Researchers have stressed the necessity of more advanced analytical methods in performance analysis for rugby union. M. T. Hughes et al. (2012), Watson, Durbach, Hendricks, and Stewart (2017) and Coughlan, Mountifield, Sharpe, and Mara (2019) highlighted that there are limitations to the univariate analysis of frequency-based PIs, and given the complex nature of rugby, there is a need for greater use of advanced analytical methods. Machine learning (ML) is a relatively new field of study that considers advanced analytical methods, and combines various disciplines including artificial intelligence, computer science, data mining and statistics. ML techniques are increasingly being applied in many disciplines including sport; however, their application to performance analysis in rugby union has only recently begun to be investigated. Recently, Bennett, Bezodis, Shearer, and

Kilduff (2020) built random forest classification models (Breiman, 2001) on PIs from the 40 matches at the group stage of the 2015 RWC, and then used these to predict win/loss outcomes for matches at the play-off stage. The authors found that 13 PIs were significant in predicting the outcome of matches at the group stage: tackle-ratio, clean breaks, average carry, lineouts won, penalties conceded, missed tackles, lineouts won in the opposition 22, defenders beaten, metres carried, kicks from hand, lineout success, penalties in opposition 22, and scrums won. Random forest models with a single variable: tackle ratio, clean breaks or average carry, were found to be able to predict 75%, 70% and 73% of matches at the group stage, respectively. A random forest model built on the group stage data could correctly predict seven out of the eight matches at the play-off stage. Clean breaks alone predicted seven out of eight matches correctly, and tackle ratio and average carry as the two PIs in the model could correctly predict six out of the eight matches. In another recent study, Watson, Hendricks, Stewart, and Durbach (2020) used convolutional and recurrent neural networks to predict the outcomes, namely territory gain, retaining possession, scoring a try, and conceding/being awarded a penalty, of sequences of play based on the order of the events and the on-field locations in which they took place.

In this study, we apply and compare a commonly used non-parametric statistical technique called the Wilcoxon signed rank test with an ML model that learns interpretable decision rules called RIPPER (W.W. Cohen, 1995) to PI data derived from group-stage and play-off stage matches at the 2019 RWC using publicly available data sourced from the official RWC website that are augmented with additional PIs that we calculate based on the original set of variables. To our knowledge, the comparison of a ML model with a statistical method is yet to be investigated in performance analysis in rugby. The interpretable nature of the decision rules generated by RIPPER is appealing and is an advantage over more black-box techniques like random forests and neural networks.

While PIs from group-stage matches at the 2015 RWC were used by Bennett et al. (2020) as inputs to construct an ML model that predicts results at the play-off stage of the tournament, our study differs in that (other than the fact we consider the 2019 RWC) we construct an ML model on both sets of tournament matches (group stage and play-off stage) but we do not use our model for prediction but rather to describe which indicators were most important for success at each stage of the tournament, and to investigate differences in PIs between winning and losing teams in matches at each of the two tournament stages. Bishop and Barnes (2013) considered the play-off stage of the 2011 tournament but not the group stage. Like van Rooyen et al. (2010) and Bennett et al. (2020), we consider play-off and group stage matches as subsets of the tournament matches. However, unlike Bennett et al. (2020), we do not use the group stage matches to predict the outcomes of the play-off matches, since important PIs at the group stage and the play-off stage can be rather different.

2. Material & Methods

2.1. Measures

The data consisting of PIs were retrieved from the official RWC 2019 website (rugbyworldcup.com). The data on the official RWC 2019 website was provided by the sports data company Stats Perform/Opta. A breakdown of the website-collected PI variables by game area is presented in Table 1.

It is often useful to express frequency-based PIs as a ratio or percentage of another PI, which can often aid interpretation in practical settings (M.D. Hughes & Bartlett, 2002). Therefore, we augmented the original set of PIs with additional calculated PIs, which are listed in Table 2. Several the attacking PIs were transformed to be based on the number of ball carries. For instance, carry metres per ball carry was calculated as a measure of carry effectiveness. Carries over the gain-line, another measure of carry effectiveness, was represented as a percentage of ball carries. Similarly, defenders beaten, line breaks and offloads were all divided by ball carries.

Out of the 45 matches in the RWC 2019 tournament, 37 were group stage matches,[1] and eight were play-off games: the final, semi-finals, quarterfinals, and bronze play-off. As mentioned, the separate consideration of the group stage and play-off stage is based on the hypothesis that there is likely to be a difference in the strategies that are effective at different stages of the tournament.

As mentioned, in this study, we take two approaches to analyse important PIs at the group and play-off stages of the tournament. The first, which we refer to as the "statistical approach" applies the Wilcoxon signed rank test. The second, which we refer to as the "ML approach" applies the decision rule algorithm RIPPER. To implement the models, it was necessary to structure the input datasets differently for each of the two approaches. For the statistical approach, the input dataset was structured such that there was one record for each of the 45 matches, and two PI variable columns for each of the 48 PI variables[2] for the winning and losing teams in each match. Thus, the dataset for the statistical approach consisted of 45 rows and a total of 96 columns. Our input dataset for the ML approach contained exactly the same information, but was structured differently such that it contained two records for each of the 45 matches in the tournament (i.e., one record each for the winning and losing teams of each match) and a column for each of the 48 PI variables, plus the won/lost class variable. Thus, the dataset for the ML approach consisted of 90 rows and 49 columns. The descriptive statistics, which are the same for both datasets, are presented along with the statistical approach results in Section 3.

2.2. Procedures

2.2.1. Statistical Approach

Since PI variables, particularly those based on frequencies, are often positively skewed, the median as well as the means of each PI are reported, along with their minimum and maximum values, and standard deviations. The descriptive statistics were generated in Microsoft Power BI.

[1] Three group-stage matches in the tournament were cancelled due to Typhoon Hagibis

[2] Because of data quality concerns, three variables related to turnovers, turnovers won, turnovers won in opposition half, and turnovers won in own half, were excluded. Turnovers won in opposition half plus turnovers won in own half were not found to add up to turnovers won in all cases. An email was sent to World Rugby regarding this; however, no response was received, thus, these three variables were excluded from the analysis.

Table 1. Performance indicators collected from the official RWC 2019 website.

Game Area	Performance Indicator
Attack	Points scored
Attack	Territory % last 10 minutes of match
Attack	Territory % whole match
Attack	Possession % whole match
Attack	Possession % first half
Attack	Carry meters
Attack	Ball carries
Attack	Ball carries over gainline
Attack	Passes made
Attack	Defenders beaten
Attack	Line breaks made
Attack	Offloads made
Breakdown	Mauls won
Breakdown	Rucks won
Kicking	Kicks from hand
Kicking	Kick meters
Kicking	Kicks regathered
Kicking	Kicks to touch
Kicking	Kicks charged down
Kicking	Kicks
Set piece	Set pieces won
Set piece	Scrums
Set piece	Scrums won
Set piece	Scrum success %
Set piece	Lineouts
Set piece	Lineouts won
Set piece	Lineout success %
Set piece	Lineout steals
Discipline	Penalties conceded
Discipline	Red cards
Discipline	Yellow cards
Defence	Tackles missed
Defence	Tackles missed
Defence	Tackle success %

A Shapiro-Wilk test was performed on each of the PI variables for winning and losing teams using RStudio (Team, 2015), and more than a third of these variables were found to be non-normally distributed. Therefore, with the relatively small sample size and the repeated measures of each teams' PIs, it was decided that the non-parametric Wilcoxon signed rank test would be used to analyse statistically significant differences between winning and losing teams. The Wilcoxon signed rank test is non-parametric in that it does not require the performance indicator variables' distributions to be normal.

The magnitude of difference is described with effect sizes (ESs) using Cohen's d (J. Cohen, 1988). While J. Cohen (1988) originally interpreted d (0.2) = small, d (0.5) = medium, d (.8) = large,

Sawilowsky (2009) provided the following rules of thumb for interpretation: d (0.01) = very small, d (0.2) = small, d (0.5) = medium, d (.8) = large, d (1.2) = very large, and d (2.0) = huge.

Table 2. Additional Performance indicators calculated based on the variables collected from the RWC website.

Game area	**Performance indicator**
Attack	Carry meters per ball carry
Attack	% of carries over gainline
Attack	Defenders beaten per ball carry
Attack	Line breaks per ball carry
Attack	Offloads per ball carry
Kicking	Average meters per kick made
Kicking	Kicks regained per kick made
Kicking	Kicks to touch per kick made
Kicking	Kicks charged per kick
Set piece	Lineout steal %
Attack	Pass to ball carry ratio %
Attack/kicking	Kick meters plus carry meters
Attack/kicking	% of meters that came from ball carries
Attack/kicking	% of meters that came from kicks

2.2.2. Machine Learning Approach

We utilised a model called RIPPER (W.W. Cohen, 1995), a decision rule algorithm that provides results in the form of rules that can be readily interpreted. Interpretability is important in sport performance analysis as it enables coaches and athletes to gain insight and identify important PIs in the hope of improving future performance. For this reason, we avoided the use of "black-box" ML algorithms such as artificial neural networks, support vector machines and random forests.

A decision rule is a simple if-then statement, which consists of a condition (antecedent) and a prediction (Molnar, 2019). Interpreting a decision rule is straightforward; to predict a new instance, start at the top and check whether the rule is applicable (i.e., the condition matches); if so, the right hand side of the rule represents the prediction for this instance (Molnar, 2019). The last rule is the default rule, which applies when none of the preceding rules have applied to an instance, thus ensuring that there is always a prediction. An advantage of RIPPER is that variable selection is performed automatically, unlike many ML models that require a priori variable selection before constructing the actual model. The utility of RIPPER for feature selection purposes in the context of match result prediction in Basketball was highlighted by Thabtah, Zhang, and Abdelhamid (2019).

Given the structure of our dataset, our problem can be treated as a classification problem in which we aim to classify teams' matches into two classes (win or loss) based on the entire set of PI variables (i.e., the original website-obtained variables in Table 1, augmented with the additional calculated variables in Table 2) for each of the group and play-off stage match datasets. The WEKA ML workbench version 3.9.3 (Hall et al., 2009) was used to construct the models. In particular, the JRip algorithm (WEKA's implementation of RIPPER) was trained on the group stage and play-off stage datasets, and was initially tuned to classify all matches correctly As shown in Figure 1, the minimum number of instances that pertain to each rule was lowered from 2 to 1, and pruning was disabled

(usePruning=False, minNo=1).[3] Note that in this study, we performed this tuning to classify all matches correctly since the model is not used for prediction and we are therefore not concerned about over-fitting. This differs from Bennett et al. (2020), whose purpose was to construct an ML model on group-stage matches that was not overly complex (over-fit) so that it was able to predict the separate set of play-off stage matches with a reasonable degree of accuracy. The purpose of our RIPPER ML model is instead to analyse separately the important PI variables at each of the two stages of the tournament but does not involve any prediction of match results.

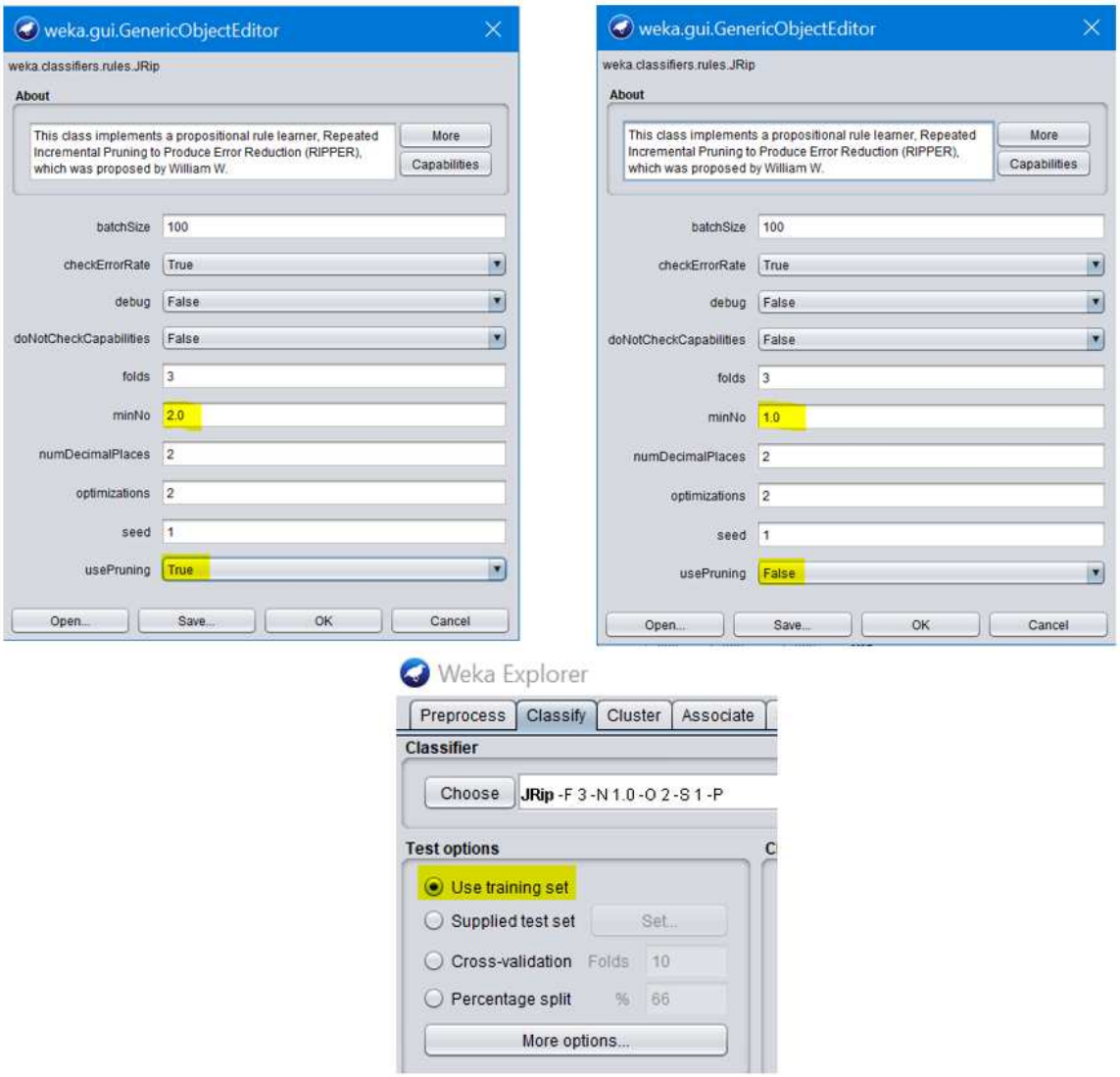


Figure 1. JRip (RIPPER) model configuration in WEKA.

[3] A J48 decision tree (WEKA's implementation of C4.5 - Quinlan, 1993), which is also readily interpretable, was also trialled, but was found to generate a much larger number of decision rules compared to RIPPER.

3. Results

The results for the group stage and play-off matches are presented. The descriptive statistics for each PI, along with the Wilcoxon signed rank test p-values and Cohen's d effect sizes are listed in Tables 3 and 4 for the group stage and play-off stage matches, respectively.

3.1. Group Stage Matches

3.1.1. Statistical Approach

Differences at the 5% level of significance are discussed in terms of median values. Where there are sets of related variables, e.g., those that relate to ball carries: carries, metres per carry, carries over the gain-line, etc., we identify the variable among this set of PIs with the highest effect size in distinguishing between winning and losing teams.

Winning teams scored 35 points, while losing teams scored 10 points ($p = 0.0000$). Carry metres had the highest effect size ($d = 0.747$) of any of the carry-related PI variables. Winning teams in group stage matches dominated carry metres with 537 metres, compared to losing teams who carried 290 metres ($p = 0.0000$). Clean breaks (in absolute terms) had a higher effect size compared to clean breaks per ball carry. Winning teams broke the opposition line 14 times, more than double that of losing teams who did so 6 times. Defenders beaten ($d = 0.660$), in absolute terms, had a higher effect size compared to defenders beaten per ball carry. Winning teams beat 32 defenders, double that of losing teams ($p = 0.0001$). Total metres gained, kick metres plus carry metres, had a statistically significant difference between winning and losing teams, with winning teams gaining 1,224 metres versus 935 metres for losing teams ($p = 0.0000$). Kicks regathered ($d = 0.678$), in absolute terms, had a higher effect size compared to kicks regathered per kick. Winning teams in group stage matches regathered 15 kicks compared to 9 for losing teams ($p = 0.0000$).

The lineout success percentage ($d = 0.424$) had a higher effect size compared to the number of lineout steals, perhaps suggesting the importance of also retaining ball on team's own throw. Winning teams had a lineout success percentage of 93.3%, compared to losing teams who had 87.5% success ($p = 0.0099$). Mauls, a common result of lineouts, particularly where teams initiate a driving maul near the opposition try line, were also significantly different between losing teams. Winning teams won five mauls compared to losing teams who won three. Winning teams also won more rucks per match. The number of offloads, in absolute terms, that teams made had a higher effect size compared to the number of offloads made per ball carry. Winning teams made 10 offloads in a match, double that of losing teams. Winning teams in group stage matches made 152 passes compared to losing teams, who made 106. Winning teams also dominated possession, in both the first half and the entire match. Similarly, winning teams had 57% of territory. Winning teams at the group stages of the tournament had higher success at scrum time, winning all their scrums as opposed to 90% of their scrums in the case of losing teams. Of the tackle-related variables, the number of tackles missed in absolute terms had the highest effect size: winning teams missed 16 tackles, half that of losing teams.

Table 3. Descriptive statistics and results for performance indicators variables for winning and losing teams across the group stage matches. Wilcoxon signed rank test p-values and Cohen's d effect sizes are reported. The far-right column indicates the sign of the difference in median values for the given performance indicator, i.e., it indicates whether winning teams had higher, lower, or equal values in the performance indicator compared to losing teams. ***, **, and * indicate statistical significance at the 10%, 5% and 1% level, respectively.

	Winning (n = 37)					Losing (n = 37)							
Performance Indicator	Average	Median	Min	Max	StdDev	Average	Median	Min	Max	StdDev	P-Value	d	
points	38.811	35	19	71	12.657	11.189	10	0	27	7.454	0.0000***	0.87	+
kick metres plus carry metres	1215.595	1224	826	1819	234.192	926.865	935	525	1403	216.831	0.0000***	0.82	+
carry metres	543.541	537	285	920	144.729	299.946	290	150	746	113.757	0.0000***	0.75	+
clean breaks	14.757	14	5	35	6.343	6.486	6	0	21	4.253	0.0000***	0.73	+
carry metres % of total metres	0.452	0.431	0.273	0.69	0.109	0.332	0.314	0.161	0.767	0.116	0.0000***	0.69	+
clean breaks per ball carry	0.111	0.108	0.032	0.203	0.041	0.063	0.062	0	0.134	0.033	0.0000***	0.69	+
kicks regathered	15.378	15	6	26	3.872	10.297	9	3	20	4.832	0.0000***	0.68	+
carries over gain-line	48.838	45	21	77	14.766	31.459	30	11	63	10.824	0.0000***	0.67	+
defenders beaten	32.324	32	7	53	11.228	17.838	16	3	48	8.512	0.0001***	0.66	+
tackles missed	17.838	16	3	48	8.512	32.297	32	7	53	11.188	0.0001***	0.66	-
carry metres per carry	4.112	4.261	1.952	5.788	0.851	3.046	2.743	1.948	5.167	0.806	0.0001***	0.64	+
carries	132.811	128	83	190	24.765	99.865	104	47	175	27.476	0.0001***	0.64	+
tackle success %	0.865	0.87	0.77	0.97	0.046	0.809	0.81	0.7	0.95	0.06	0.0001***	0.63	+
passes made	160.514	152	95	254	40.461	114.568	106	40	253	43.151	0.0003***	0.60	+
possession first half	0.568	0.55	0.37	0.79	0.095	0.432	0.45	0.21	0.63	0.095	0.0003***	0.59	+
defenders beaten per ball carry	0.244	0.229	0.064	0.411	0.077	0.181	0.17	0.035	0.364	0.07	0.0004***	0.59	+
offloads	9.838	10	2	24	4.705	5.486	5	0	21	3.839	0.0004***	0.58	+
kicks regathered per kick	0.54	0.5	0.261	1.071	0.167	0.379	0.333	0.103	1.429	0.224	0.0006***	0.56	+
percentage of carries over gain-line	0.361	0.358	0.253	0.446	0.054	0.315	0.324	0.182	0.444	0.064	0.0019***	0.51	+
territory	0.573	0.57	0.37	0.81	0.122	0.427	0.43	0.19	0.63	0.122	0.0029***	0.49	+
scrum success %	0.936	1	0	1	0.175	0.866	0.9	0	1	0.197	0.0070***	0.44	+
offloads made per ball carry	0.073	0.073	0.016	0.136	0.03	0.052	0.047	0	0.135	0.026	0.0072***	0.44	+
lineout success %	0.922	0.933	0.667	1	0.083	0.851	0.875	0.5	1	0.118	0.0099***	0.42	+
mauls won	5.054	5	0	10	2.66	3.324	3	0	7	2.08	0.0107**	0.42	+
possession	0.545	0.54	0.37	0.76	0.094	0.455	0.46	0.24	0.63	0.094	0.0111**	0.42	+
tackles made	114.324	114	47	183	33.123	138.757	129	72	218	34.687	0.0133**	0.41	-
lineout steals	1.243	1	0	5	1.303	0.541	0	0	2	0.72	0.0231**	0.37	+
rucks won	85.541	80	46	127	19.699	72.405	73	27	122	22.436	0.0381**	0.34	+
penalties conceded	7.514	8	3	15	2.637	9	9	3	17	3.425	0.0516	0.32	-
scrums	6.189	6	0	13	2.749	7.595	7	0	13	3.192	0.0623*	0.31	-
pass to ball carry ratio	1.205	1.234	0.779	1.458	0.165	1.129	1.098	0.743	1.5	0.194	0.0726	0.3	+
red cards	0.027	0	0	1	0.162	0.162	0	0	1	0.369	0.0726	0.3	=
kicks charged per kick	0.016	0	0	0.065	0.02	0.028	0	0	0.167	0.04	0.1134	0.26	=
lineout steal %	0.099	0.071	0	0.5	0.105	0.066	0	0	0.5	0.107	0.1287	0.25	+
kick metres	672.054	690	267	1028	194.13	626.919	621	223	1073	208.293	0.1698	0.23	+
kick metres % of total metres	0.548	0.569	0.31	0.727	0.109	0.668	0.686	0.233	0.839	0.116	0.1698	0.23	-
scrums won	5.973	6	0	13	2.746	6.919	7	0	13	3.123	0.1743	0.22	-
kicks from hand	23.838	25	9	39	7.023	22.216	23	9	39	5.946	0.1951	0.21	+
kicks	29.757	31	12	47	7.084	28.054	28	14	42	6.089	0.2095	0.21	+
lineouts won	11.892	13	4	19	3.812	10.703	11	4	15	3.118	0.2934	0.17	+
yellow cards	0.27	0	0	2	0.643	0.405	0	0	2	0.676	0.3977	0.14	=
kicks charged	0.514	0	0	2	0.683	0.703	0	0	3	0.926	0.4142	0.13	=
set pieces won	18.108	18	6	26	4.572	17.568	17	12	25	3.538	0.5437	0.10	+
kicks to touch per kick	0.389	0.409	0.194	0.696	0.119	0.399	0.385	0.154	0.64	0.113	0.5831	0.09	+
kick metres per kick	22.565	23.226	14.74	29.03	3.875	22.009	22.429	12.39	32.16	4.313	0.5975	0.09	+
territory last 10 minutes	0.496	0.42	0.1	0.99	0.244	0.504	0.58	0.01	0.9	0.244	0.8504	0.03	-
lineouts	12.73	13	5	21	3.775	12.541	12	6	20	3.326	0.8775	0.03	+
kicks to touch	11.27	11	5	22	3.71	11.054	11	4	21	3.518	0.9122	0.02	=

3.1.2. Machine Learning Approach

According to the decision rules generated by RIPPER (Figure 2), losing teams at the group stage had low carry metres (less than or equal to 343 metres) and a lineout success percentage less than or equal to 93.3%. This conjunctive rule explained 26 out of the 37 (70%) losing teams' matches at the group stage. Teams that had a high number of missed tackles (33 or more) despite having carry metres of 341 or more, also lost, explaining six out of 37 (16%) of the losing teams' matches at the group stage. Four out of the 37 (11%) losing teams' matches were the result of low carry effectiveness and few clean breaks per carry. In Fiji's unexpectedly loss to Uruguay, Fiji had very low kicking metres (227 kicking metres, while the median for losing teams at the group stage was 621 metres).

The rules generated by RIPPER not only automatically identify the important features but also provide an idea of their importance through calculating the percentage of losing (or winning) matches that fall under a given rule. For instance, since 70% of losing teams' matches fell within the first rule generated by RIPPER, we can observe that carry metres and lineout success percentage were, jointly, the most important factors at the group stage of the tournament.

```
JRIP rules:
===========

(carry_metres <= 343) and (lineout_success_% <= 0.933) => result=lost (26.0/0.0)
(tackles_missed >= 33) and (carry_metres >= 341) => result=lost (6.0/0.0)
(clean_breaks_per_ball_carry <= 0.074534) and (percentage_of_carries_over_gainline <= 0.335404) => result=lost (4.0/0.0)
(kick_metres <= 227) => result=lost (1.0/0.0)
 => result=won (37.0/0.0)
```

Figure 2. RIPPER-generated rules – group stage matches.

3.2. Play-off Stages

3.2.1. Statistical Approach

The differences in PIs between winning and losing teams in play-off matches at the 5% and 10% levels of significance are described as are non-statistically significant differences with effect sizes exceeding 0.5

Winning teams (mdn = 29 points) in play-off matches typically scored nearly double the points of losing teams (mdn = 15 points) (p =0.014). On attack, winning teams (mdn = 96.5) made significantly fewer ball carries per match than losing teams (mdn = 131.5) (p = 0.04). Winning teams (mdn = 108.5) also made significantly fewer passes per match than losing teams (mdn = 154.5) (p =0.04), which resulted in no significant difference in the pass to ball carry ratio between winning and losing teams. On defence, interestingly, winning teams (mdn = 150) made significantly more tackles per match than losing teams (mdn = 105) (p = 0.04).

Winning teams had higher total metres from kick metres plus carry metres (mdn = 1126.5 metres) compared to losing teams (mdn = 987.5 metres) (p = 0.08). Winning teams (mdn = 45%) had lower possession (mdn = 55%) than losing teams (p = 0.09). In addition, winning teams won fewer rucks (mdn = 71.5) compared to losing teams (mdn = 96) (p = 0.08).[4] Winning teams (mdn = 35.2%) had a greater percentage of their carries over the gain-line compared to losing teams (mdn = 29.6%) (p = 0.08). In terms of the kicking area of the game, winning teams kicked more from hand (mdn = 29.5 kicks) compared to losing teams (mdn = 22 kicks) (p = 0.05). In addition, winning teams did not have any of their kicks charged, while losing teams had (a median of) one kick charged (p = 0.05); 3.9% of losing teams kicks were charged down (p = 0.06).

Winning teams had superior ball carry effectiveness (mdn = 3.6 metres per carry), gaining more metres per ball carry than losing teams (mdn = 2.2 metres per carry) (p = 0.14, d = 0.52). Winning teams (mdn = 784 metres) also had higher kicking metres (mdn = 565.5 metres) (p = 0.14, d = 0.52). On defence, winning teams (mdn = 90%) had a slightly higher tackle success percentage than losing teams (mdn = 86.5%) (p = 0.12, d = 0.55).

[4] Unfortunately, the RWC official website did not contain data related to ruck success percentages, dominant tackles, or statistics by field zone, e.g., 22-halfway etc, which would have allowed for a more in-depth analysis.

Table 4. Descriptive statistics and results for performance indicators variables for winning and losing teams across the play-off matches. Wilcoxon signed rank test p-values and Cohen's d effect sizes are reported. The far-right column indicates the sign of the difference in median values for the given performance indicator, i.e., it indicates whether winning teams had higher, lower, or equal values in the performance indicator compared to losing teams. ***, **, and * indicate statistical significance at the 10%, 5% and 1% level, respectively.

	Winning (n = 8)					Losing (n = 8)							
Performance Indicator	**Average**	**Median**	**Min**	**Max**	**StdDev**	**Average**	**Median**	**Min**	**Max**	**StdDev**	**p-value**	**d**	
Points	30.25	29	19	46	10.109	13	15	3	19	5.099	0.0141**	0.87	+
Carries	106.125	96.5	71	156	30.498	138.875	131.5	114	181	20.763	0.0423**	0.72	-
passes made	122.25	108.5	67	185	41.4	161.75	154.5	115	211	36.752	0.0423**	0.72	-
tackles made	161.5	150	145	206	22.344	111.75	105	74	164	28.195	0.0421**	0.72	+
kicks charged	0.125	0	0	1	0.331	1.125	1	0	3	1.053	0.0545*	0.68	-
kicks charged per kick	0.003	0	0	0.025	0.008	0.034	0.039	0	0.073	0.029	0.0591*	0.67	-
kicks from hand	28	29.5	18	37	5.979	24.25	22	15	36	7.241	0.0680*	0.65	+
kick metres plus carry metres	1155.25	1126.5	920	1515	193.937	1021.875	987.5	682	1661	307.073	0.0801*	0.62	+
percentage of carries over gain-line	0.354	0.352	0.278	0.435	0.05	0.298	0.296	0.202	0.411	0.059	0.0801*	0.62	+
rucks won	74.625	71.5	48	110	19.937	100.75	96	87	137	15.658	0.0801*	0.62	-
possession	0.448	0.45	0.36	0.56	0.062	0.553	0.55	0.44	0.64	0.062	0.0898*	0.6	-
tackle success %	0.889	0.9	0.81	0.93	0.038	0.849	0.865	0.77	0.89	0.039	0.1226	0.55	+
carry metres per carry	3.683	3.576	2.762	4.754	0.641	2.623	2.2	1.418	4.149	1.014	0.1415	0.52	+
kick metres	773.625	784	505	1182	203.208	647.5	565.5	391	1163	249.557	0.1415	0.52	+
kicks	32	33.5	23	40	5.831	29.125	27	17	41	8.146	0.1422	0.52	+
yellow cards	0.375	0	0	1	0.484	0	0	0	0	0	0.1489	0.51	=
clean breaks per ball carry	0.095	0.095	0.041	0.172	0.039	0.059	0.058	0.016	0.119	0.034	0.1834	0.47	+
defenders beaten per ball carry	0.187	0.175	0.135	0.279	0.045	0.144	0.135	0.096	0.221	0.041	0.1834	0.47	+
possession first half	0.436	0.415	0.32	0.62	0.083	0.564	0.585	0.38	0.68	0.083	0.2033	0.45	-
kicks to touch per kick	0.29	0.302	0.217	0.346	0.043	0.369	0.378	0.195	0.529	0.119	0.262	0.4	-
offloads	6.125	4.5	2	14	4.484	9.75	11.5	2	15	4.63	0.271	0.39	-
offloads made per ball carry	0.052	0.042	0.022	0.098	0.028	0.069	0.08	0.018	0.107	0.032	0.2936	0.37	-
lineout success %	0.87	0.882	0.615	1	0.113	0.941	0.967	0.818	1	0.066	0.3096	0.36	-
territory	0.464	0.44	0.38	0.62	0.083	0.536	0.56	0.38	0.62	0.083	0.3456	0.33	-
kick metres per kick	24.248	22.697	18.38	31.95	4.847	22.063	22.777	15.69	28.37	3.976	0.3627	0.32	-
mauls won	4.5	3.5	1	10	2.784	2.875	1.5	0	7	2.803	0.4017	0.3	+
carries over gain-line	38.25	33.5	23	64	14.411	41.75	42.5	23	58	11.155	0.4461	0.27	-
kicks to touch	9.375	9.5	5	12	2.446	10.125	11	5	14	2.571	0.479	0.25	-
scrums won	6.375	6	3	11	2.395	4.875	4	3	8	1.763	0.5513	0.21	+
kicks regathered per kick	0.519	0.466	0.189	1	0.223	0.581	0.538	0.242	1.095	0.251	0.6241	0.17	-
lineouts	10.375	9.5	6	20	4.27	11.5	12	7	15	2.828	0.6215	0.17	-
territory last 10 minutes	0.539	0.525	0.22	0.92	0.21	0.46	0.47	0.08	0.78	0.21	0.6241	0.17	+
scrums	6.625	6.5	3	11	2.497	5.125	4	3	9	1.9	0.5513	0.1	+
set pieces won	15.875	16	13	21	2.368	15.125	15.5	10	22	3.822	0.7998	0.09	+
carry metres % of total metres	0.336	0.319	0.22	0.535	0.093	0.366	0.355	0.164	0.596	0.132	0.8336	0.07	-
clean breaks	10.125	7.5	5	21	5.644	8.375	7	2	16	4.872	0.8334	0.07	+
kick metres % of total metres	0.664	0.681	0.465	0.78	0.093	0.634	0.645	0.404	0.836	0.132	0.8334	0.07	+
lineout steal %	0.074	0	0	0.308	0.109	0.072	0.056	0	0.2	0.076	0.8339	0.07	-
pass to ball carry ratio	1.138	1.168	0.944	1.268	0.104	1.163	1.109	0.888	1.581	0.203	0.8336	0.07	+
defenders beaten	19.875	16.5	12	34	7.928	20.375	20.5	11	34	7.193	0.8885	0.05	-
kicks regathered	16.25	15	7	26	6.629	15.75	14.5	8	24	5.309	0.8884	0.05	+
tackles missed	20.375	20.5	11	34	7.193	19.875	16.5	12	34	7.928	0.8885	0.05	+
penalties conceded	8.5	8	6	13	2.121	8.5	8	5	12	2.236	0.9322	0.03	=
carry metres	381.625	356.5	275	580	99.896	374.375	343	173	639	171.081	0.9442	0.02	+
lineouts won	9.5	9	5	18	3.841	10.25	9	7	15	3.031	0.9438	0.02	=
lineout steals	0.875	0	0	4	1.364	0.625	0.5	0	2	0.696	1	-	-
red cards	0	0	0	0	0	0.125	0	0	1	0.331	1	-	=
scrum success %	0.968	1	0.857	1	0.056	0.955	1	0.75	1	0.086	1	-	=

3.2.2. Machine Learning Approach

When applied to the play-off matches, the RIPPER-generated rules (Figure 3) showed that winning teams won 78 rucks or less. This was the most important factor at this stage of the tournament, accounting for six out of eight (75%) of the winning teams' matches at the play-off stage. On the other hand, despite winning more than 78 rucks, in two play-off matches (New Zealand in their quarterfinal win against Ireland and England in their semi-final win against New Zealand) the winning teams made 55 or more carries over the gain-line. The territory in the last 10 minutes of the match variable appears to be somewhat less relevant, and indeed, when we manually removed this variable from the set of PIs, only the Wales-New Zealand bronze-play-off match was misclassified (Figure 4).

```
JRIP rules:
===========

(rucks_won <= 78) => result=won (6.0/0.0)
(carries_over_gainline >= 55) and (territory_last_10_mins <= 0.39) => result=won (2.0/0.0)
 => result=lost (8.0/0.0)

Number of Rules : 3
```

Figure 3. RIPPER-generated rules – play-off stage matches.

```
JRIP rules:
===========
(rucks_won <= 78) => result=won (6.0/0.0)
(carries_over_gainline >= 55) => result=won (2.0/1.0)
=> result=lost (7.0/0.0)

Number of Rules : 3
```

Figure 4. RIPPER-generated rules for the play off matches when the component of the rule related to the territory in the last 10 minutes (territory_last_10_mins <= 0.39) is removed, which is seemingly irrelevant. In this case, Wales losing to New Zealand is incorrectly classified as a win by the "(carries_over_gainline >= 55) => result=won" rule (In this match, the bronze play-off, Wales won 137 rucks and had 58 carries over the gain-line, but lost the match).

4. Discussion

As expected, there were differences in the performance indicators that contributed to success at the play-off stage compared to the group stage, which suggests the need for teams to adjust their playing strategies at the play-off stage in order to be successful.

At both the group and play-off stages of the tournament, effective ball carries, as measured by the percentage of ball carries that penetrated the opposition gain-line, as well as total metres gained (kick metres plus carry metres), were found to contribute to success. On the other hand, while dominating possession, carrying the ball more frequently, making more passes, winning more rucks, and making less tackles contributed to success at the group stage of the tournament, the opposite was the case at the play-off stage.

At the group stage of the tournament, in comparing our statistical results to those of Bennett et al. (2020), who studied the 2015 RWC, tackle success, clean breaks, average carry (metres per carry), missed tackles, defenders beaten, carry meters, lineout success were found to be important PIs at the group stages of the 2015 as well as the 2019 tournament. While Bennett et al. (2020) found that lineouts won, penalties conceded, kicks from hand and scrums won were also important at the group stage at the group stage of the 2015 RWC, we found in our statistical results that these PIs were not important in distinguishing successful and unsuccessful teams at the group stage of the 2019 RWC (kicks regathered and scrum success were found to be important at the group stage of the 2019 RWC, however).[5]

At the play-off stage of the tournament, the results of the statistical approach found that winning teams made less carries, made less passes, and won less rucks compared to losing teams. Bennett et al. (2020) found that tackle success, clean breaks, average carry in metres, missed tackles, defenders beaten, carry metres, lineout success were important PIs at the group stage of the 2015 RWC, results that are consistent with those of the present study. We also found that there were no statistically significant differences in the ball-to-carry ratios or total carry meters between winning and losing teams at the play-off stage. Despite being in possession of the ball a lower percentage of the time, winning teams in play-off matches were more effective in their carries in terms of both the percentage of their carries that penetrated the gain-line, as well as in metres gained per ball carry. Winning teams made more kicks out of hand and gained more metres via this tactic, resulting in higher total metres gained through either kicks or ball carries. Winning teams also pressured and occasionally charged down opposition kicks, while losing teams were unable to do so. Interestingly, winning teams made more tackles per match, suggesting spending time on defence was not necessarily detrimental, provided their defence was solid, with winning teams having a slightly higher tackle success percentage. Bishop and Barnes (2013), in studying the play-off stages of the 2011 RWC, found that winning teams played more of a territory and kicking style of game rather than a possession-based game. Although our results did not show a significant difference in territory between winning and losing teams, they do suggest that the 2019 RWC was similar in that a possession-based/pick-and-go type of strategy was not effective at the play-off stage of the tournament. Unlike A. Hughes et al. (2017) who studied the 2015 RWC, we did not find that the percentage of opposition lineouts stolen discriminated between winning and losing teams at the play-off stage of the 2019 tournament.

The most important factors identified at the play-off stage by the ML approach were rucks won and number of carries over the gain-line, which were both also identified as important through the statistical approach. Carry metres being an important factor at the group stage is consistent with the

[5] Bennett et al. (2020) also found that penalties and lineouts won in the opposition 22m zone were important at the group stage of the 2015 RWC, however these PIs were unfortunately not available on the 2019 RWC official website.

results obtained via the statistical approach, with carry metres found to have the largest effect size (apart from points). However, the decision rules approach also identified a joint relationship in which low carry metres together in conjunction with a low lineout success percentage was the most important factor contributing to losing at the group stage of the tournament, explaining 70% of losing teams' matches at the group stage (lineout success did not have an overly large effect size in the results from the Wilcoxon signed rank test, although it was still significant at the 1% level). A high number of tackles missed, making few clean breaks per carry, and having low carry effectiveness also contributed to losses at the group stage. This again suggests that a possession-based game with a repeated pick-and-go type strategy may not have been effective at the play-off stage of the 2019 RWC, which agrees with the results obtained via the statistical approach.

The results of the ML approach again highlight that forming a large number of rucks was not an advantage at the play-off stage of the tournament, but the two teams that did have a large number could otherwise win through carry effectiveness in terms of having a high number of carries that penetrated the gain-line.

5. Conclusion

Our statistical and ML approaches provided somewhat different results, most notably in the number of important PIs identified, with RIPPER selecting a small subset of the original PIs, perhaps a disadvantage since it allows for a less in-depth analysis. The obvious advantages of decision rules are that, compared to statistical approach, which required the calculation of many Wilcoxon signed rank tests for each of the PI variables, the decision rules are fast and easy to generate and interpret. Like non-parametric statistical tests like the Wilcoxon signed rank test, decision rules do not require distributional assumptions such as normality. On the other hand, a weakness of decision rules is that, particularly for a small number of matches (e.g., we only have eight play-off matches), some rules generated may be relatively random in nature, but happen to classify the match outcomes correctly (evident in the "territory in the last 10 minutes" variable appearing in the RIPPER-generated rules for the play-off matches).

The present study is not without limitations. The variables included as performance indicators were limited to those that were available on the 2019 RWC official website. Performance indicators such as dominant tackles, and ruck frequency, ruck success percentage or rucks lost were not available. In addition, variables were not available by field position, e.g., 22 metre line to halfway, 22 metre line to try line, etc. This limited the ability to compare with the results of some prior studies. Also, only team-level performance indicators were considered in the present study, player-level variables were not.

An interesting avenue for future work could be to augment performance indicator variables with external variables such as venue, weather, referees and so on, and conduct a comparative analysis of their importance. Another would be to experiment with other interpretable machine learning models on other similar datasets that consist of performance indicator variables.

References

Bennett, M., Bezodis, N. E., Shearer, D. A., & Kilduff, L. P. (2020). Predicting performance at the group-phase and knockout-phase of the 2015 rugby world cup. *European Journal of Sport Science*, 1–9.

Bishop, L., & Barnes, A. (2013). Performance indicators that discriminate winning and losing in the knockout stages of the 2011 rugby world cup. *International Journal of Performance Analysis in Sport*, *13* (1), 149–159.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences, 2nd ed. Hillsdale, NJ: Erlbaum.

Cohen, W. W. (1995). Fast effective rule induction. In *Machine learning proceedings* 1995 (pp. 115-123). Morgan Kaufmann.

Coughlan, M., Mountifield, C., Sharpe, S., & Mara, J. K. (2019). How they scored the tries: applying cluster analysis to identify playing patterns that lead to tries in super rugby. *International Journal of Performance Analysis in Sport*, 1–17.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.

Hughes, A., Barnes, A., Churchill, S. M., & Stone, J. A. (2017). Performance indicators that discriminate winning and losing in elite men's and women's rugby union. *International Journal of Performance Analysis in Sport, 17* (4), 534–544.

Hughes, M., & White, P. (1997). An analysis of forward play in the 1991 rugby union world cup for men. *Notational analysis of sport I & II*, 183–191.

Hughes, M. D., & Bartlett, R. M. (2002). The use of performance indicators in performance analysis. *Journal of sports sciences, 20* (10), 739–754.

Hughes, M. T., Hughes, M. D., Williams, J., James, N., Vučković, G., & Locke, D. (2012). Performance indicators in rugby union. *Journal of Human Sport & Exercise*.

Hunter, P., & O'Donoghue, P. (2001). A match analysis of the 1999 rugby union world cup. In Books of abstracts fifth world congress of performance analysis in sports (pp. 85–90).

McCorry, M., Saunders, E., O'Donoghue, P., & Murphy, M. (2001). A match analysis of the knockout stages of the 1995 rugby union world cup. *Notational analysis of sport III*. Cardiff: UWIC, 230–239.

Molnar, C. (2019). Interpretable machine learning. *Lulu. com*.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods, 8*(2), 26.

Stanhope, J., & Hughes, M. (1997). An analysis of scoring in the 1991 rugby union world cup for men. *Notational Analysis of Sport I y II*, 167–176.

Team, R. O. (2015). RStudio: integrated development for r. *RStudio, Inc., Boston, MA URL http://www.rstudio.com, 42*, 14.

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA game result prediction using feature analysis and machine learning. *Annals of Data Science, 6*(1), 103-116.

van Rooyen, K. M., Diedrick, E., & Noakes, D. T. (2010). Ruck frequency as a predictor of success in the 2007 rugby world cup tournament. *International Journal of Performance Analysis in Sport, 10* (1), 33–46.

van Rooyen, K. M., Lambert, I. M., & Noakes, D. T. (2006). A retrospective analysis of the IRB statistics and video analysis of match play to explain the performance of four teams in the 2003 rugby world cup. *International Journal of Performance Analysis in Sport, 6* (1), 57–72.

van Rooyen, K. M., & Noakes, D. T. (2006). Movement time as a predictor of success in the 2003 rugby world cup tournament. *International journal of Performance analysis in Sport, 6* (1), 30–39.

Vaz, L., Vasilica, I., Kraak, W., & Arrones, S. L. (2015). Comparison of scoring profile and game related statistics of the two finalists during the different stages of the 2011 rugby world cup. *International Journal of Performance Analysis in Sport, 15* (3), 967–982.

Watson, N., Durbach, I., Hendricks, S., & Stewart, T. (2017). On the validity of team performance indicators in rugby union. International Journal of Performance Analysis in Sport, 17 (4), 609–621.

Watson, N., Hendricks, S., Stewart, T., & Durbach, I. (2020). Integrating machine learning and decision support in tactical decision-making in rugby union. *Journal of the Operational Research Society*, 1-12.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin 1*, 6 (1945).

CHAPTER 7: PUBLICATION 4 - "THE APPLICATION OF MACHINE LEARNING TECHNIQUES FOR PREDICTING MATCH RESULTS IN TEAM SPORT: A REVIEW"

This survey paper considered performance at the match outcome level. A wide range of studies published between 1996 and 2019 that used machine learning to predict match results in team sports were critically analysed and synthesised. The number of papers published in this domain expanded greatly in the 2010s. Specifically, the study considered team sports, including invasion sports (e.g., soccer, rugby union, and basketball) and striking/fielding sports, such as baseball and cricket. Commonly applied machine learning models, data preprocessing methodologies, and model evaluation approaches were discussed in detail, and the findings were synthesised to produce domain-related insights. The predictive accuracies achieved across different sports and the characteristics of some sports that make them inherently more difficult to predict than others were also discussed in detail.

In terms of practical implications, the study identified a general lack of available benchmark datasets available for sports match result prediction, and those that do exist, such as the Open International Soccer Database (Dubitzky et al., 2019), only contain goals scored as features derived from in-play events. The paper also highlighted that the differences between sports, datasets, features, and evaluation metrics make inter-study comparisons of results difficult.

However, it was noted that model performance could be evaluated in other ways against baselines and other heuristics or predictions, for example, simple rule-based predictions (e.g., always predicting that the home team will win), random guesses, predictions derived from betting odds, and predictions by human experts. It was mentioned that betting odds are, in fact, an excellent predictor of match outcome and are, therefore, a helpful benchmark to try and outperform with machine learning (Wilkens, 2021; Tax & Joustra, 2015) (bookmaker consensus models that aggregate the odds of several bookmakers as per Leitner, Zeileis, & Hornik, 2010 can also be used as a benchmark).

The survey also emphasised that although neural networks were commonly applied in early studies in the domain (e.g., Purucker, 1996 and Kahn, 2003), a better approach is to select a set of candidate models based on a survey of the literature and conduct a comparative

evaluation of their performances based on an appropriately selected evaluation metric. Selecting and engineering a relevant set of features for a model was more critical for achieving high predictive accuracy than having a large dataset regarding the number of matches available. Indeed, we found that over 80% of the papers surveyed suggested engineering richer and more informative features as an avenue for future work. This is consistent with machine learning in general, whereby domain knowledge to generate more informative and discriminatory features is generally considered the best strategy to improve predictive accuracy (Domingos, 2012).

The significance of this work in terms of its contribution to the literature was in providing a comprehensive overview of machine learning for team sports outcome prediction up to 2019, providing insights into the methodologies used, challenges faced, and potential future directions in this field. The study aimed to be a resource for researchers and practitioners interested in applying machine learning to sports match result prediction. For selecting machine learning model features, a greater degree of interdisciplinary collaboration between sports performance analysis researchers and machine learning researchers was identified as necessary for driving future research innovation in this domain.

Journal of Artificial Intelligence Research 73 (2022) 1285-1322 Submitted 11/2021; published 04/2022

The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review

Rory Bunker RORY.BUNKER@G.SP.M.IS.NAGOYA-U.AC.JP
Graduate School of Informatics
Nagoya University
Furocho, Chikusa Ward
Nagoya, Aichi 464-8601, Japan

Teo Susnjak T.SUSNJAK@MASSEY.AC.NZ
School of Mathematical and Computational Sciences
Massey University
Massey University East Precinct Albany Expressway, SH17, Albany
Auckland 0632, New Zealand

Abstract

Predicting the results of matches in sport is a challenging and interesting task. In this paper, we review a selection of studies from 1996 to 2019 that used machine learning for predicting match results in team sport. Considering both invasion sports and striking/fielding sports, we discuss commonly applied machine learning algorithms, as well as common approaches related to data and evaluation. Our study considers accuracies that have been achieved across different sports, and explores whether evidence exists to support the notion that outcomes of some sports may be inherently more difficult to predict. We also uncover common themes of future research directions and propose recommendations for future researchers. Although there remains a lack of benchmark datasets (apart from in soccer), and the differences between sports, datasets and features makes between-study comparisons difficult, as we discuss, it is possible to evaluate accuracy performance in other ways. Artificial Neural Networks were commonly applied in early studies, however, our findings suggest that a range of models should instead be compared. Selecting and engineering an appropriate feature set appears to be more important than having a large number of instances. For feature selection, we see potential for greater inter-disciplinary collaboration between sport performance analysis, a sub-discipline of sport science, and machine learning.

1. Introduction

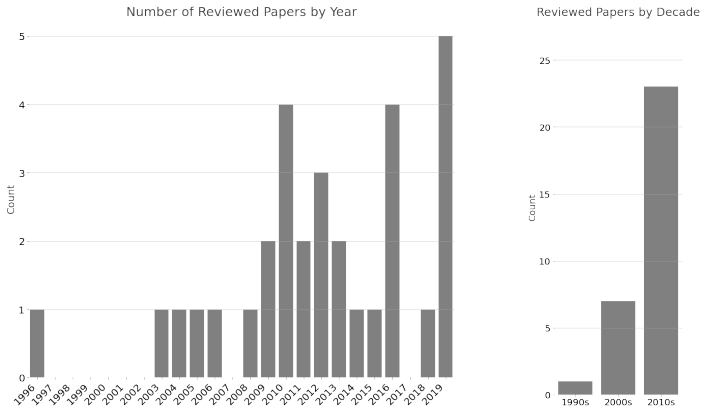
Sport result prediction is an interesting and challenging problem due to the inherently unpredictable nature of sport, and the seemingly endless number of potential factors that can affect results. Indeed, it is this unpredictable nature that is one of the main reasons that people enjoy sport. Despite its difficulty, predicting the results of sports matches is of significant interest to many different stakeholders, including bookmakers, bettors, and fans. Interest in match result prediction has grown with the increased availability of sport-related data online and with the emergence of online sports betting. Sport experts and former players often make predictions on upcoming matches, which are commonly published in the media.

The application of machine learning (ML) in sport has recently branched widely, and research surveys exist on predicting sports injuries (Van Eetvelde et al., 2021), applying predictive analytics to sport training (Rajšp & Fister, 2020), and using predictions for determining optimal team formations (Ishi & Patil, 2021). Numerous studies also exist that focus on in-play predictions, which predict the occurrence of specific events during a contest, e.g., a goal being scored by a specific player. However, the focus of the present survey is specifically on the application of ML in predicting the outcomes of matches in team sports. The prediction of final match results is of interest not only to the likes of bookmakers and bettors but also to players, team management and performance analysts in order to identify the most important factors in achieving winning outcomes. One important task within outcome prediction in sport is to select the best set of features for the predictive model. If the model used is interpretable and incorporates some feature selection mechanisms or, alternatively, a feature selection method is applied prior to applying the model, the most important predictive features can be extracted. Although some features, e.g., the match venue, officials, weather, etc., are external to the sport match, in-play features may identify areas in which teams can adjust their tactics/strategy to improve performance. Sport performance analysis, a sub-discipline of sport science, considers performance indicators (PIs), which are a selection or combination of action variables that aim to define some or all aspects of a performance, which, in order to to be useful, should relate to a successful performance or outcome (Hughes & Bartlett, 2002). In this context, the match outcome is, of course, one measure of performance. Therefore, important features (PIs) are useful for players, team management and sport performance analysts to identify ways in which they can improve their performance and achieve winning outcomes. As we will discuss further below, in this domain we see potential for greater collaboration between sport performance analysis researchers and machine learning researchers going forward.

In the academic literature, match result prediction in sport has been considered by the statistics and operations research communities for some time; however, the application of ML techniques for this purpose is more recent. The first study in this domain appears to have been published in 1996 (Purucker, 1996); however, it was not until the 2010s when research activity in this area intensified, as shown in Figure 1(b). This review considers research papers in this field over the last three decades, but places an inclusion constraint requiring that papers must have made use of at least one ML technique. To that end, knowledge-based systems (e.g., fuzzy logic- and rule-based systems), and ratings methodologies (e.g., Elo ratings) (Rotshtein et al., 2005; Tsakonas et al., 2002; Min et al., 2008) were not considered to be in-scope. Due to the large number of papers that have been published recently, we needed to constrain the scope of this review further to only focus on the prediction of *match results* in *team sports*. Thus, we did not include studies related to sports played by individuals or in pairs, e.g., horse racing (Davoodi & Khanteymoori, 2010), swimming (Edelmann-Nusser et al., 2002), golf (Wiseman, 2016), tennis (Somboonphokkaphan & Phimoltares, 2009) and javelin (Maszczyk et al., 2014). Although this review has a narrower scope compared to some prior reviews (Table 1), its contribution lies in providing a more in-depth analysis of the application of ML for sport result prediction in team sports than the previous survey articles. The present review introduces the reader to a broad overview of approaches in which ML techniques have been applied for match result prediction in team sports. A distinct contribution of this work, in comparison to prior

reviews, lies in our discussion of how the characteristics of particular team sports potentially play a role in the ability of ML to be able to accurately predict match results. For instance, we explore whether the predictability of matches may depend on whether it is an invasion (time-dependent) sport or a striking/fielding (innings-dependent) sport. We also discuss how the occurrence and increments of points or goals may also affect the predictability of match results. Furthermore, we comment on what some of the key drivers of successful studies have been with respect to how ML has been applied, and how those elements have contributed to higher predictive accuracies. In addition, this study makes a meaningful contribution in identifying future research trends and opportunities in this field, while combining the findings into a useful set of recommendations for other professionals in this domain. Lastly, as mentioned, we also highlight the possibility for greater collaboration between researchers from sport performance analysis and those from machine learning. The lack of collaboration between these fields has often resulted in unclear and inconsistent terminology, and there are significant opportunities to advance research with a more unified approach.

The remainder of this paper is structured as follows. In Section 2, we outline our methodological approach and inclusion criteria for this study. Then, in Section 3, we review the literature in ML for match result prediction in team sport, categorizing sports by type (invasion sports and striking/fielding sports) and further sub-categorizing studies based on the individual sport. We also present tabular summaries of the results of the surveyed studies by sport. Following this, we provide critical analysis and discussion in Section 4, before concluding in Section 5.



(a) Papers by year

(b) Papers by decade

Figure 1: Figures depicting the number of papers surveyed by (a) year and (b) decade.

2. Methodology

In this section, we firstly outline our inclusion criteria, which specifies the scope of this review. Secondly, we categorize the algorithms into groups in order to identify trends in usage over time. Thirdly, we describe the measure of accuracy, which has been (by far) the most widely-used metric of performance in this application domain, and we also discuss

why it is indeed an appropriate metric for match result prediction in team sport. Finally, we describe recurring future research themes extracted from the surveyed studies.

2.1 Inclusion Criteria

We considered studies that used ML to predict match results in team sports, which were published between 1996 and 2019 (inclusive). To be included in this review, a study must have applied at least one ML technique. Thus, studies that only considered fuzzy logic-, ratings- (e.g., Elo) or statistical-based methods were not included.

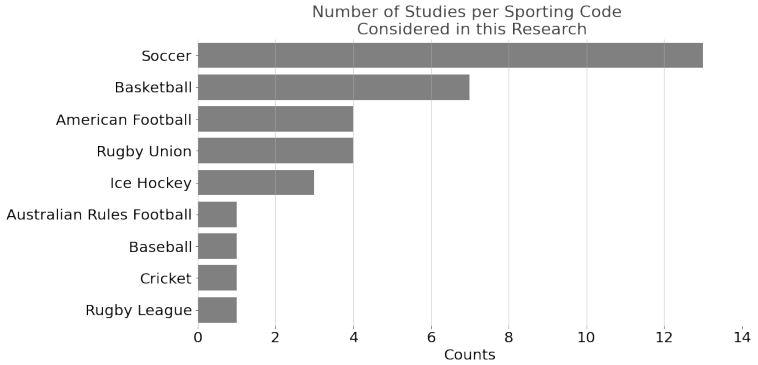


Figure 2: Number of surveyed studies by sporting code

A total of 31 papers were considered in this review. The distribution of papers by year and decade of publication can be seen in Figure 1, which shows an explosion in the level of interest in this field from 2009 onward. The 31 papers covered nine distinct sports and, given that some papers studied multiple sports, there were a total of 35 sport studies. The distribution of the number of studies by sport is shown in Figure 2.

We identified four related review papers that were published in the same time period that we considered. Table 1 shows the methods applied, the types of sports considered, and what the applied methods covered were aiming to achieve. Taken together, we consider these aspects to define the scopes of the prior survey articles. Both Haghighat et al. (2013) and Keshtkar Langaroudi and Yamaghani (2019) included studies that used ML as well as knowledge-based systems (e.g., systems based on fuzzy and rule-based logic). Haghighat et al. (2013) covered both team and non-team sports, focusing on match result prediction specifically, while Keshtkar Langaroudi and Yamaghani (2019) also included studies that considered tactical decision making. The review of Razali et al. (2018) had a narrower scope compared to other reviews (and compared to the present review), and focused on studies applying one specific ML model, a Bayesian Network (BN), for predicting the match results of one sport, soccer. The survey of Beal et al. (2019) had a wide scope, considering artificial intelligence methods (of which machine learning, data mining, and knowledge-based methods are sub-categories) for match result prediction, as well as tactical decision making, player investments, and injury prediction. Beal et al. (2019) also considered fantasy sports in addition to team sports.

Unlike prior work, our review provides more up-to-date coverage and draws out more critical insights compared to previous studies, e.g., Haghighat et al. (2013), who did not, for

Survey	Methods	Sports	Purpose
Haghighat et al. (2013)	Machine Learning & Knowledge-based Systems	Team Sports & Non-Team Sports	Match result prediction
Razali et al. (2018)	Bayesian Networks	Soccer	Match result prediction
Beal et al. (2019)	Artificial Intelligence	Team Sports & Fantasy Sports	Match result prediction, tactical decision making, player investments, injury prediction
Keshtkar Langaroudi and Yamaghani (2019)	Machine Learning & Knowledge-based Systems	Team & Non-Team Sports	Match result prediction, tactical behavior
This review	Machine Learning	Team Sports	Match result prediction

Table 1: Comparison of the scope of this review with that of similar surveys covering a similar time period

instance, discuss the characteristics of the sports that they considered, and how these may have had an impact on obtained accuracies. They also did not consider whether accuracies obtained have improved over time, nor did they discuss the appropriateness of different models for the purpose of sport result prediction, e.g., in terms of their interpretability.

For this paper, we sought to define the scope of the survey such that it is sufficiently narrow so that the analysis is adequately in-depth but, at the same time, its scope is not too wide such that the number of surveyed papers is unmanageable. In this review, we focus on match result prediction, which is a team-level measure of performance. The scope of Keshtkar Langaroudi and Yamaghani (2019) was somewhat ill-defined in that they also included, e.g., coverage of Tilp and Schrapf (2015), who considered the use of ANNs for analyzing tactical defensive behavior in handball, which is not related to the prediction of player- or team-level results. Although narrower in scope, our review provides a wider coverage of surveyed papers than Keshtkar Langaroudi and Yamaghani (2019) and Haghighat et al. (2013), surveying more than double and triple the number of studies, respectively. In addition, unlike the above-mentioned review papers, we cover the studies that arose from the 2017 Open International Soccer Database Competition, which was important in terms of creating a benchmark dataset for soccer match result prediction (the review of Haghighat et al., 2013 noted that there were no such benchmark datasets at the time of their review).

2.2 Algorithm Grouping

ML algorithms were grouped into families of algorithms to identify trends in their usage patterns. All variants of ANN, such as Back Propagation (BP), Feed-Forward, as well as Self-Organising Maps and Long Short-Term Memory (LSTM) ANNs, were grouped under the same umbrella of algorithms. CART, C4.5 and WEKA's J48 algorithm were grouped under the Decision Tree family of methods. RIPPER, FURIA and ZeroR were merged into the Rule Sets category, while the Local Weighted Learning (LWL) algorithm was merged with k-Nearest-Neighbors (kNN). Additionally, AdaBoost, XGBoost, LogitBoost, RobustBoost, and RDN-Boost were grouped in the Boosting category, while Sequential Minimal Optimization (SMO) (Kohonen, 1990) was combined with Support Vector Machines (SVMs). Methods that combined several different families of ML algorithms into a single decision-making architecture were termed Ensemble. Although Naïve Bayes and BNs share some common theoretical foundations, a decision was made to keep them separate since the latter includes the ability to incorporate domain knowledge to a larger degree, which was a

motivating factor for its usage in some studies. A histogram depicting the usage patterns of these algorithm groups, sorted according to their frequency of use, can be seen in Figure 3.

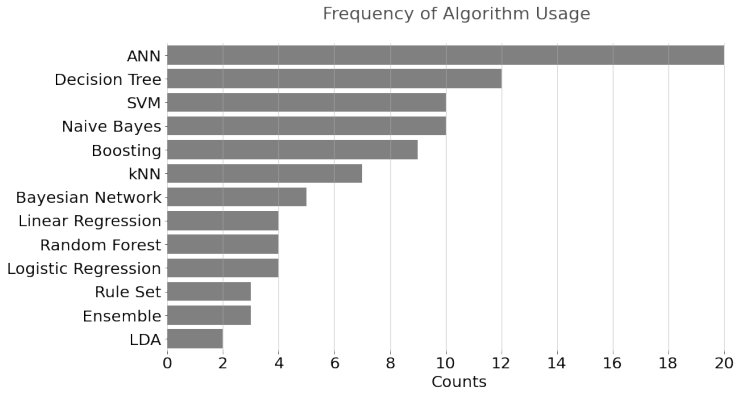


Figure 3: Histogram of usage patterns of the algorithm groups covered in this survey

2.3 Performance Evaluation Metrics

To allow for meaningful between-study comparisons and interpretations, we consider the accuracy measure as the primary evaluation metric, which the vast majority of surveyed studies used. Accuracy is defined as the number of correct predictions divided by the total number of predictions or, equivalently, the proportion of the sum of total true positive and true negative classifications divided by all samples in the dataset. Although the charts presented in the remainder of this paper only include studies that reported accuracy as the measure of performance, all results, including one paper that reported balanced accuracy and three that reported the average ranked probability score (RPS), are presented in the results summary tables in Section 3. It should be noted that the results from binary-class and multi-class problems are not directly comparable since a given level of accuracy becomes harder to achieve as the number of classes increases. Thus, the accuracies of studies that used a three-class formulation instead of a two-class formulation are excluded from the charts for comparative purposes but, again, are reported in the summary tables.

2.4 Future Research Themes

In order to draw out insights across all papers in terms of what the general future research direction trends might be, a set of recurring general themes were first extracted from all of the papers. Subsequently, all of the text referring to future research directions across all of the papers was encoded based on the extracted themes, and a histogram was rendered that depicts the frequency of each theme.

3. Literature Review

In this section, we review the studies identified by our inclusion criteria, distinguishing between invasion sports and striking/fielding sports (Figure 4). Subsection 3.1 covers studies

related to invasion sports, including American Football, Rugby Union, Soccer, Basketball and Ice Hockey, and Subsection 3.2 considers the striking/fielding sports of Baseball and Cricket. The results of the surveyed studies are also summarized in tables containing the competition, models applied, the number of matches in the original dataset, as well as the best performing model and the number of features used in that model.

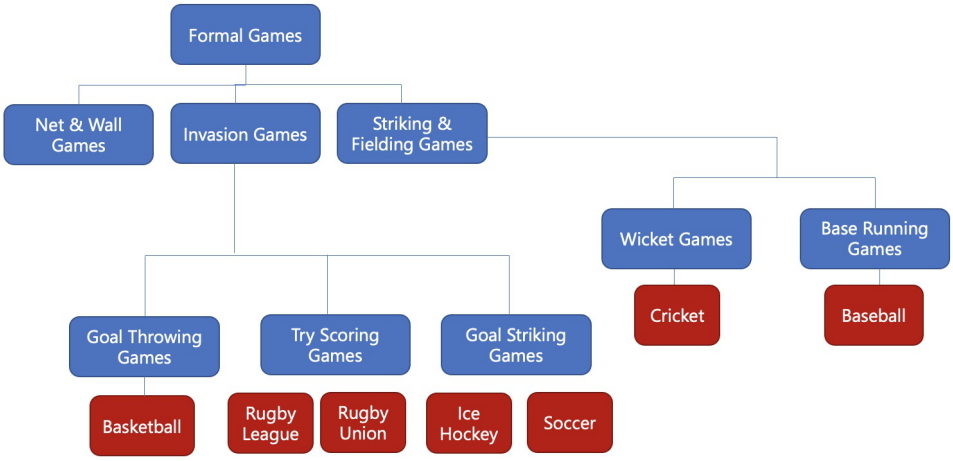


Figure 4: Classification of formal games in this survey based on the categorizations provided by Read and Edwards (1992) and Hughes and Bartlett (2002).

3.1 Invasion Sports

Invasion sports are time dependent in that they have matches with fixed periods of time, commonly divided into halves or quarters. The aim for teams in invasion sports is to move into the opposition team's territory by maintaining possession, creating space, and attacking a goal or target in order to convert scoring opportunities into points while, at the same time, defending their own space and goal to prevent the opposition from scoring (Mitchell, 1996).

3.1.1 American Football

Purucker (1996) used an ANN and unsupervised learning techniques to predict the results of National Football League (NFL) football matches, using data from 90 matches from weeks 11 to 16 of the 1994 NFL competition. Six features were included: victories, yardage differential, rushing yardage differential, turnover margin, time in possession, and betting odds (the inclusion of betting line odds improved upon initial results). An ANN trained with BP was used, with BP providing the best performance among the different network training methods. The unsupervised learning methods that were applied were: the Hamming error (Hamming, 1950), Adaptive Resonance Theory (ART) (Carpenter & Grossberg, 2003), and Self-Organizing Maps (SOM). The SOM provided the best performance among the unsupervised methods, however, it could not match the performance of the ANN. Matches from weeks 12 to 15 were used to predict week 16 matches, with the ANN correctly predicting 11 out of the 14 matches (78.6%) in week 16. Weeks 12 to 14 were also used to predict week

15, with the ANN correctly predicting 10 of the 14 matches (71.4%) in week 15. The author recognized that the dataset consisted of a small number of matches and features, and mentioned that improvements could be gained by fine-tuning the encoding, ANN architecture, and training methods.

Kahn (2003) predicted NFL matches using data from 208 matches in the 2003 season. A BP-trained ANN was used, and the features included were: total yardage differential, rushing yardage differential, time in possession differential, turnover differential, a home or away indicator, home team outcome, and away team outcome. The numeric features were calculated based on the 3-week historical average (the feature's average value over the past 3 weeks), as well as its average value over the entire season. Using the average over the entire season achieved higher accuracy. Weeks 1 to 13 of the 2003 competition were used as training data, and weeks 14 and 15 as the test set. Accuracy of 75% was achieved, which was slightly better than expert predictions on the same matches. It was suggested that, in future work, betting odds and team rankings could be included as features, and matches from previous seasons could be used for model training.

David et al. (2011) used a committees-of-committees approach with ANNs to predict NFL matches, where many networks were trained on different random partitions of the data. For training, 500 ANNs were used, and the best 100 were used in each committee, of which 50 were ultimately used. The mean was used to determine the vote of each committee and to combine their predictions. The features used were differentials between the home and away teams based on: passing yards, yards per rush play, points, interceptions, and fumbles. The differentials between the home and away teams were to incorporate the well-known home advantage phenomenon. The season average of these features were used, apart from the first five rounds of the season, where the weighted average between the current season's and previous season's features was used. In particular, 100% of the feature value from the previous season was used for week 1, then 80% of the previous season and 20% of the current season in week 2, and so on until week 6, at which point only the current season value was used. A total of 11 inputs were used in the ANN for each game, and the Levenber-Marquadt (LM) routine was used to train the network. Principal Components (PCA) and derivative based analyses were applied to determine which features were most influential. It was found that 99.7% of the variance in their data was due to the passing and rushing differentials. The results were compared to bookmakers and predictions from thepredictiontracker.com, and were found to be comparable to the predictions of bookmakers, and better than most of the online predictions. The avenues for future work were to use different types of ANN, e.g., RBF, to include additional statistics (e.g., possession, strength-of-schedule, kicking game and injuries), to investigate how to best predict games that are early in the season, and to apply the method to other levels of football (e.g., NCAA) as well as to other sports.

Delen et al. (2012) used an SVM, CART Decision Tree and ANN to predict the results of NCAA Bowl American Football matches. The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework (Wirth & Hipp, 2000) was used as the experimental approach, and the dataset contained 28 features and 244 matches. A classification approach that predicted home win/away win was compared with a numeric prediction approach that predicted points margin (home team points minus away team points). CART provided the best performance, with 86% accuracy (using 10-fold cross validation), which was statistically significantly better than the other models. The classification approach produced

Paper	Models used	No. of features	No. of Matches	Accuracy of best model
Purucker (1996)	ANN trained with BP*, Unsupervised: Hamming, ART, SOM	6	90	75% (week 14 and 15 combined)
Kahn (2003)	ANN trained with BP	10	208	75%
David et al. (2011)	Committees of ANNs trained with LM	11	unknown	unknown
Delen et al. (2012)	SVM, CART*, ANN	28	244	86%

Table 2: American Football Studies (* denotes the best performing model).

better results than numeric prediction. When the models were trained on the 2002/2003 to 2009/2010 seasons and then tested on the 2010-2011 season, CART again had the best performance, achieving 82.9% accuracy. The suggested directions for future research were to include more variables or represent them in different forms, to make use of other classification and regression methods (e.g., rough sets, genetic algorithm based classifiers, ensemble models), to experiment with seasonal game predictions by combining static and time series variables, and to apply the approach to other sports.

3.1.2 Rugby Union

O'Donoghue and Williams (2004) compared the predictive ability of human experts with computer-based methods for 2003 Rugby World Cup matches. Multiple Linear Regression, Logistic Regression, an ANN, and a Simulation Model were the computer-based models used. Data from the previous four world cups were used to develop predictive models of match results based on three variables: team strength, determined by synthesising world rankings (actual world rankings had not yet been introduced at the time of the 2003 Rugby World Cup), distance travelled to the tournament as a measure of home advantage, and the number of recovery days between matches. The computer-based models correctly predicted between 39 and 44.5 of the 48 matches, while the 42 experts correctly predicted an average of 40.7 matches. However, there was far greater variation in the accuracies of the experts, with the most successful person correctly predicting 46 of the 48 matches. The most accurate computer-based model was the Simulation Model.

O'Donoghue et al. (2016) compared the accuracy of 12 Linear Regression models in predicting match results at the 2015 Rugby World Cup. The models differed in terms of: (i) whether or not the assumptions of the predictive modeling technique were satisfied or violated, (ii) whether all (1987-2011) or only recent Rugby World Cup tournaments' (2003-2011) data were used, (iii) whether the models combined pool and knockout stage match data, and (iv) whether the models included a variable that tried to capture a relative home advantage. The common independent variable in all models was the relative quality, which was the difference from the higher ranked team's world ranking points. The dependent variable was the points margin. All models were executed 10,000 times within a simulation package that introduced random variability. The best model achieved accuracy of 74% and, notably, match outcomes in international Rugby appeared to be more difficult to predict

Paper	Competition	Models used	No. of features	No. of Matches	Accuracy of best model
O'Donoghue and Williams (2004)	2003 Rugby World Cup	Multiple linear regression, Logistic Regression, ANN, Simulation model*	3	48	93%
Reed and O'Donoghue (2005)	English Premiership Rugby	Multiple linear regression, ANN*, discriminant function analysis	7	498	46.1%
McCabe and Trevathan (2008)	Super Rugby	ANN trained with BP* and CGD	19	unknown	74.3%
O'Donoghue et al. (2016)	2015 Rugby World Cup	Linear regression*, Simulation	1	280	74.4%

Table 3: Rugby Union Studies (*denotes the best performing model).

than in previous years. The best model used data from all previous Rugby World Cups in a way that violated the assumptions of Linear Regression, using only one independent variable and ignoring the relative home advantage, while generating separate models for the pool and knockout stage matches.

3.1.3 Soccer

Joseph et al. (2006) found that incorporating expert knowledge into a BN model can result in strong performance, especially when the sample size is small. A decision tree (MC4) and kNN model were also used to predict the results of Soccer matches played by the English Premier League (EPL) team, Tottenham. Their dataset consisted of 76 matches. Four variables were included in the expert model, while 30 variables were used in their original, general model. The expert BN was found to provide the best performance, achieving 59.2% accuracy when predicting a home win, away win or draw (a 3-class problem). In the future, the authors proposed to develop a more symmetrical BN model using similar data but for all the teams in the EPL, to incorporate player-quality features (e.g., players that have performed at international level), and to add additional nodes such as attack quality and defence quality.

Buursma (2010) used data from 15 years of Dutch Soccer to predict match results, and was interested in: which variables were important for predicting match results, how the probabilities for the match results can be predicted from these variables, and what bets should be placed to maximize profit. The author applied the following models in WEKA: Classification via Regression, Multi-class Classifier, Rotation Forest, Logit Boost, BN, Naïve Bayes, and ZeroR. There were three match outcomes to be predicted: home win, draw, and away win. The feature set consisted of 11 features, and all features were either aggregated or averaged across a team's previous 20 matches (by experimentation, 20 was found to be the best number of matches to average across). Classification via Regression and the Multi-class Classifier had the best prediction accuracy, both achieving 55%. For future work, the author considered including more features, e.g., yellow/red cards, the number of players each team has, the management, player budgets and home ground capacities, and was also interested in applying the model to other similar sports such as basketball, baseball and ice hockey.

Huang and Chang (2010) used a BP-trained ANN to predict the results of the 64 matches in the 2006 Soccer World Cup tournament. The output of the ANN was the relative ratio between two teams for each match, which could then be converted into an odds ratio. Eight features were selected based on the domain knowledge of the authors: goals for, shots,

shots on goal, corner kicks, direct free kicks on goal, indirect free kicks on goal, possession, and fouls conceded. Accuracy of 76.9% was achieved based on 2-class prediction, i.e., not including draws. The ANN was found to have difficulty predicting a draw, which was a frequent outcome in the group stages.

In contrast to Joseph et al. (2006), Hucaljuk and Rakipović (2011) found that incorporating expert opinion did not produce any improvement for soccer match result prediction. Their dataset consisted of 96 matches (6 rounds) in the regular part of the European Champions League competition, and was split into three different training and test datasets: a 3 round-3 round training-test split, a 4 round-2 round split, and a 5 round-1 round split. Feature selection resulted in 20 features in their basic feature set. An additional feature set consisted of the 20 basic features plus variables selected by experts. Naïve Bayes, BNs, Logit Boost, kNN, a Random Forest, and a BP-trained ANN were compared, with the ANN performing the best, achieving accuracy of 68%. Perhaps surprisingly, the expert-selected features were not found to yield any improvement. It was mentioned that further improvements could be gained by refining the feature selection, modeling player form, and obtaining a larger dataset.

Odachowski and Grekow (2012) analysed fluctuations in betting odds and used ML techniques to investigate the value in using such data for predicting soccer matches. The odds for home win, away win and draw for the preceding 10 hours, measured at 10-minute intervals, were tracked over time (the data was obtained from the Betfair and Pinnacle Sports websites). A total of 32 features were computed from this time-series, e.g., maximum and minimum changes in betting odds, overall changes in odds, and standard deviations. The 10-hour period was divided into thirds and the features were also calculated based on these sampling periods. The authors balanced their dataset such that there were an equal number of home wins, away wins, and draws (372 matches in each class). Six classification algorithms from WEKA were compared: BN, SMO, LWL, Ensemble Selection, Simple CART, and a Decision Table, and feature selection methods in WEKA were applied. It was found that draws were especially difficult to correctly predict, with only around 46% accuracy obtained when attempting 3-class classification. However, accuracy of around 70% was obtained when ignoring draws. Discretization and feature selection methods were found to improve results. The authors suggested that additional features describing changes in betting odds could be included in future work.

Tax and Joustra (2015) aimed to identify easily retrievable features that have utility for predicting soccer match results. In particular, they sought to explore whether there were different levels of utility between features broadly classified as public match data sources and those derived from bookmaker odds. Their experiments covered data from the 2000 to 2013 seasons of the Dutch Eredivisie competition. The researchers conducted numerous experiments comparing the Naïve Bayes, LogitBoost, ANN, Random Forest, CHIRP, FURIA, DTNB, J48 and Hyper Pipes algorithms from the WEKA toolkit. The experiments were performed in conjunction with feature selection methods such as PCA, Sequential Forward Selection and ReliefF. The best results on the public match data sources were achieved by the Naïve Bayes and ANN classifiers in combination with PCA, achieving accuracy of 54.7%. FURIA achieved the highest accuracy using bookmaker odds features with 55.3%; however, this was ultimately not found to be statistically significant. A marginal improvement in accuracy was realized with LogitBoost and ReliefF when both bookmaker odds and public

match-data features were used, producing an accuracy of 56.1%. While this was also not at a statistically significant level, it nonetheless pointed to the potential utility in combining a broader variety of features for further investigation.

Prasetio (2016) used Logistic Regression to predict EPL soccer results in the 2015/2016 season. Their data set consisted of six seasons, from the 2010/2011 to the 2015/2016 seasons (2,280 matches). Home offense, away offense, home defence, and away defence were used as the input features, and it was found that home defence and away defence were significant (the authors did not mention how these offense and defence ratings were constructed). Despite this, the model that included all four variables was found to yield higher accuracy. Four different training-test splits were trialed, producing four different sets of model coefficients. The best performing model achieved 69.5% accuracy based on a 2-class problem (excluding draws). In the future, they remarked that the results could be used to assist management with game strategy, or the trained models could be turned into a recommendation system for evaluating player purchase decisions.

Danisik et al. (2018) applied a Long Short-Term Memory (LSTM) neural network (Hochreiter & Schmidhuber, 1997) to predict match results in a number of soccer leagues. Classification, numeric prediction and dense approaches were compared, and were contrasted with an average random guess, bookmakers' predictions, and the most common class outcome (home win). Player-level data were included, which was obtained from FIFA video games. Incorporating player-level and match history data, a total of 139 features were included (134 in the dense model). Four EPL seasons, 2011/2012 to 2015/2016, were considered, comprising a total of 1,520 matches. The average accuracy obtained for a 3-class classification problem was 52.5%, achieved with the LSTM Regression model, using the 2011/2012, 2012/2013 and 2015/2016 seasons as the training dataset and the 2013/2014 season as the validation dataset. The accuracy obtained for a two-class problem (excluding draws) was 70.2%. It was stated that betting odds and additional match-specific player and team features could be included, and the use of convolution to transform input attributes during training and a deeper exploration of the ability of LSTMs to leverage features like specific tactics could be investigated.

The Open International Soccer Database: Prediction Challenge. In 2017, a significant development took place for ML researchers interested in Soccer. A comprehensive, open-source database called the Open International Soccer Database (Dubitzky et al., 2019) was compiled and made public. The database contains over 216,000 matches from 52 leagues and 35 countries. The motivation behind this project was to encourage ML research in Soccer by building an up-to-date knowledge base that can be used on an ongoing basis for the prediction of real-world soccer match outcomes, as well as to act as a benchmark dataset that makes comparisons between experiments more robust. In order to maximize the utility of this database, a deliberate design choice was made to collect and integrate only data that are readily available for most soccer leagues worldwide, including lower leagues. The consequence of this is that the database lacks fields that are highly-specialized and sophisticated. Subsequent to its creation, the 2017 Soccer Prediction Challenge was conducted, and a competition was held based on this dataset, the results of which were published in a special issue of the *Machine Learning* (Springer) Journal. The challenge involved building a single model to predict 206 future match results from 26 different Soccer leagues, which were to be played between March 31 and April 9, 2017. Unlike most prior studies, which used accuracy

as a performance metric, this competition used the average ranked probability score (RPS), which measures how good forecasts are compared to observed outcomes when the forecasts are expressed as probability distributions. In the remainder of this section, we summarize three of the ML-related papers from this competition.

Hubáček et al. (2019b) experimented with both relational- and feature-based methods to learn predictive models from the database. Pi-ratings (Constantinou et al., 2012), which capture both the current form and historical strengths of teams, and a rating based on PageRank (Page et al., 1999) were computed for each team. XGBoost (regression and classification) algorithms were employed as the feature-based method, and RDN-Boost was used as the relational method. The feature-based classification method with XGBoost performed best on both the validation set and the unseen challenge test set, achieving 52.4% accuracy on the test set. Avenues for future work were to augment the feature set, to weight aggregated data by recency, to include expert guidance in forming relational concepts (e.g., using active learning), and to identify features that are conditionally important given occurrences of certain features at higher levels of the tree.

Constantinou (2019) created a model combining dynamic ratings with a Hybrid BN. The rating system was partly based on the pi-rating system of Constantinou and Fenton (2013), computing a rating that captures the strength of a team relative to other teams in a competition. Pi-ratings update based on how the actual goal differences in a match compare to the expected result (based on existing ratings). The rating calculation involves placing more emphasis on the result rather than the margin, so the effect of large goal differences are dampened. Unlike the original pi-ratings, this version also incorporated a team form factor, searching for continued over- or under-performance. Four ratings features (two for the home and away teams) were used as the BN inputs. The model provided empirical evidence that a model can make good predictions for a match between two specific teams, even when the prediction was based on historical match data that involved neither of those two teams. The model achieved accuracy of 51.5% on the challenge test data set. The author recognized the limited nature of this data set, and mentioned that incorporating other key factors or expert knowledge, e.g., player transfers, key player availability, international competition participation, management, injuries, attack/defence ratings, and team motivation/psychology may be beneficial.

Berrar et al. (2019) developed two methods, with two feature sets for result prediction: recency and rating features. Recency feature extraction involved calculating the averages of features over the previous nine matches, based on four feature groups: attacking strength, defensive strength, home advantage and opposition strength. Rating features were based on the performance ratings of each team, and were updated after each match based on the expected and observed match results and the pre-match ratings of each team. The XGBoost and kNN algorithms were applied to each of these two feature sets, with both performing better on the rating feature set. The best performance overall was obtained with XGBoost on the rating features, although this result was achieved post-competition. Their best model that was submitted for the competition was a kNN that was applied to the rating features, which achieved accuracy of 51.9% on the unseen competition test set. It was mentioned that the generally small number of goals per match and narrow margins of victory in soccer meant that it is difficult to make predictions based on goals only. The authors concluded that innovative feature engineering approaches and effective

incorporation of domain knowledge are critical for sport result prediction modeling, and are likely to be more important than the choice of ML algorithm. It was again recognized that the competition dataset was limited, and that data about various game events (e.g., yellow and red cards, fouls, possession, passing and running rates, etc.), players (e.g., income, age, physical condition) and teams or team components (e.g., average height, attack running rate) would likely help to improve results.

Overall, the Open International Soccer Database competition produced a number of innovative methods and approaches. Notably, researchers commonly combined some form of ratings-based method with ML techniques. Despite having access to a very large number of matches available in the competition dataset, all of the studies found that accuracy levelled off after a certain point, which perhaps indicates that having a broad range of predictive features is critical for predicting match results in sport. As mentioned, a significant weakness of the competition dataset is that it does not contain in-play features that occur during matches.

3.1.4 Basketball

Loeffelholz et al. (2009) predicted National Basketball Association (NBA) match results in the 2007/2008 season. There were 620 matches that were used for training and testing, and 30 that were used as the validation set, treated as "un-played" games. The features included were: field goal percentage, three-point percentage, free-throw percentage, offensive rebounds, defensive rebounds, assists, steals, blocks, turnovers, personal fouls, and points. To predict the un-played games, averages of features in the current season were found to result in better performance than averaging the features across the past five matches. The authors also investigated ANN fusion using Bayesian Belief Networks and Neural Networks. Four different types of ANN were applied: a Feed-Forward NN (FFNN), a Radial Basis Function NN (RBF-NN), a Probabilistic NN (PNN) and a Generalized NN (GRNN). The best model, a FFNN with four shooting variables, correctly predicted the winning team 74.3% of the time on average, which was better than USA Today, who achieved 68.7%. An iterative Signal-to-Noise Ratio (SNR) method was used for feature selection, selecting four out of the 22 original variables. Although fusion did not result in higher accuracy on this dataset, the authors mentioned that it still warranted further investigation. They also mentioned that different features could be used in the baseline model, and the models could be adjusted to determine whether they can beat the betting odds rather than only predicting the winning team.

Zdravevski and Kulakov (2009) obtained two seasons of NBA data (1,230 matches), the first of which was used as the training dataset, and the second of which was used as the test dataset. All of the algorithms in the WEKA machine learning toolkit were applied with their default parameter settings. A set of 10 features was selected by the authors. Classification accuracy of 72.8% was achieved with Logistic Regression. It was stated that, in future work, it would be preferable to compare their predictions to those of experts, and that it might be possible to cluster training and test data to use different models on each cluster in order to account for winning and losing streaks. It was also mentioned that aggregations or ensembles of classifiers (e.g., voting schemes) could be investigated, and that

Paper	Competition	Models used	No. of features	No. of Matches	Accuracy of best model
Reed and O'Donoghue (2005)	English Premier League	Multiple Linear Regression, ANN*, Discriminant Function Analysis	7	498	57.9% (3-class)
Joseph et al. (2006)	English Premier League	Bayesian Network, Expert Bayesian Network*, Decision Tree, kNN	4	76	59.2% (3-class)
McCabe and Trevathan (2008)	English Premier League	ANN trained with BP* and CGD	19	unknown	54.6% (3-class)
Buursma (2010)	Dutch Eredivisie League	WEKA: MultiClassClassifier with ClassificationViaRegression*, RotationForest, LogitBoost, Bayesian Network, Naïve Bayes, ZeroR	11	4590	55% (3-class)
Huang and Chang (2010)	2006 Soccer World Cup	ANN trained with BP	8	64	62.5% (3-class), 76.9% (2-class)
Hucaljuk and Rakipović (2011)	European Champions League	Naïve Bayes, Bayesian network, LogitBoost, kNN, random forest, ANN*	20	96	68% (3-class)
Odachowski and Grekow (2012)	Various leagues	BayesNet*, SVM, LWL, Ensemble Selection, CART, Decision Table	320	1,116	70.3% (2-class), 46% (3-class)
Tax and Joustra (2015)	Dutch Eredivisie League	WEKA: NaïveBayes, LogitBoost*, ANN, RandomForest, CHIRP, FURIA, DTNB, J48, HyperPipes	5	4284	56.1% (3-class)
Prasetio (2016)	English Premier League	Logistic Regression	4	2280	69.5% (2-class)
Danisik et al. (2018)	Various leagues	LSTM NN classification, LSTM NN regression*, Dense Model	139	1520	52.5% (3-class), 70.2% (2-class)
Hubáček et al. (2019b)	52 leagues	XGBoost classification*, XGBoost regression, RDN-Boost	66	216,743	52.4% (3-class)
Constantinou (2019)	52 leagues	Hybrid Bayesian Network	4	216,743	51.5% (3-class)
Berrar et al. (2019)	52 leagues	XGBoost*, kNN	8	216,743	51.9%** (3-class)

Table 4: Soccer Studies (*denotes the best performing model). Accuracies for 2-class (win,loss) and 3-class (win,loss,draw) problems are denoted. **Berrar et al. (2019)'s best performing model was kNN, but post-competition they mention that they improved on this with XGBoost. The accuracy of 51.9% is for their in-competition result - the XGBoost accuracy was not reported in their paper but it would have been slightly higher than this.

automatic feature selection methods should be used rather than features being manually selected by the authors.

Ivanković et al. (2010) used a BP-trained ANN to predict the results of the Serbian First B Basketball League, using five seasons from 2005/2006 to 2009/2010 (890 matches). The authors used the CRISP-DM framework for their experimental approach, and investigated how the successful shot percentage in six different regions of the court affected match results. The input dataset was divided into training and testing data (75%:25%), and 66.4% accuracy was obtained on the test set. The authors then reverted back to the data preparation phase of the CRISP-DM framework to see whether adding additional variables (offensive rebounds, defensive rebounds, assists, steals, turnovers, and blocks) could improve results. This improved accuracy to just under 81%. It was concluded that actions in the zone directly under the hoop, in particular, rebounds in defence and scoring in this zone, were crucial to determining the outcome of the game. It was mentioned that, in future work, a richer data set and new software solutions may help to ensure that all relevant events are included.

Miljković et al. (2010) predicted basketball match results using data from 778 games in the regular 2009/2010 NBA season. The features were divided into game (in-play) features, which directly relate to events within the match (e.g., fouls per game and turnovers per game), and those that relate to standings (e.g., total wins and winning streaks). Naïve Bayes achieved 67% accuracy (10-fold cross validation), and was found to be the best performing model when compared to kNN, a Decision Tree and SVM. Their future research plans included applying their system to other sports, and to experiment with other models such as ANNs.

Cao (2012) created an automated data collection system, obtaining six years of NBA matches from the 2005/2006 season to the 2010/2011 season. The dataset, comprising around 4,000 matches, was divided into training, test, and validation sets. Four models were compared: Simple Logistic Regression, Naïve Bayes, SVM, and a BP-trained ANN. The feature set of 46 variables was selected based on the domain knowledge of the author. All models were found to produce similar accuracies, with Simple Logistic Regression, which performs automatic feature selection, achieving the highest accuracy (67.8%). The best expert predictions on teamrankings.com were slightly better, achieving 69.6% accuracy. The author suggested that, in the future, clustering could be used to group players by positional group, or to identify outstanding players, and outlier detection methods could be used to identify outstanding players or team status. Investigating the impact of player performance on match results, and comparing different feature sets derived from box-scores and team statistics were also mentioned as avenues for future work.

Shi et al. (2013) investigated the viability of ML in the context of predicting the results of individual NCAAB matches which, up to this point, had been dominated by statistical methods. They used the WEKA toolkit to compare the accuracies of the ANN, C4.5, RIPPER and Random Forest algorithms. Experiments were conducted using data from six seasons, together with an expanding window approach so that initially training was performed on the 2008 season and testing on the 2009 season. Thereafter, the combination of all previous seasons comprised the training set until the 2013 season. The authors concluded that, on average, the ANN with default parameter settings provided the best accuracies, though statistical tests to confirm this were not provided. The top-ranked features in

Paper	Competition	Models used	No. of features	No. of Matches	Accuracy of best model
Loeffelholz et al. (2009)	NBA	ANN (types: FFNN*, RBG, PNN, GRNN, fusions of these)	4	650	74.3%
Zdravevski and Kulakov (2009)	NBA	All models in WEKA (Logistic Regression*)	10	1,230	72.8%
Ivanković et al. (2010)	Serbian First B	ANN trained with BP	51	890	81%
Miljković et al. (2010)	NBA	kNN, Decision Tree, SVM, Naïve Bayes*	32	778	67%
Cao (2012)	NBA	Simple Logistic Regression*, Naïve Bayes, SVM, ANN	46	4,000	67.8%
Shi et al. (2013)	NCAAB	ANN*, C4.5 Decision Tree, RIPPER, Random Forest	7	32,236	74%
Thabtah et al. (2019)	NBA	ANN, Naïve Bayes, LMT Decision Tree*	8	430	83%

Table 5: Basketball Studies (*denotes the best performing model).

terms of importance were location, the "four factors" (Oliver, 2002) and adjusted offensive and defensive efficiencies (kenpom.com). The authors remarked that they experienced an upper limit of 74% accuracy that they could not improve beyond, and noted that feature engineering and selection hold promise for an improvement in results.

Thabtah et al. (2019) used Naïve Bayes, an ANN, and an LMT Decision Tree model to predict the results of NBA matches, focusing on trialing various different subsets of features in order to find the optimal subset. Their dataset, obtained from Kaggle.com, consisted of 21 features and 430 NBA finals matches from 1980 to 2017 and a binary win/loss class variable. Defensive rebounds were found to be the most important factor influencing match results. The feature selection methods used were: Multiple Regression, Correlation Feature Subset (CFS) selection (Hall, 1998), and RIPPER (Cohen, 1995). Defensive rebounds were selected as being important by all three feature selection methods. The best performing model (83% accuracy) was trained on a feature set consisting of eight features, which were selected with RIPPER and trained using the LMT model. The authors suggested that the use of a larger dataset, more features (e.g., players, coaches), and other models (e.g., function-based techniques and deep learning methods) are potential avenues for further research.

3.1.5 Ice Hockey

Weissbock et al. (2013) noted that Ice Hockey had not received much attention in ML research with respect to predicting match results. The continuous nature of this sport makes it difficult to analyze, due to a paucity of quantifiable events such as goals. This characteristic was cited as a possible reason for the lack of attention Hockey had received historically. The authors focused on exploring the role of different types of features in predictive accuracies across several types of ML algorithms. They considered both traditional statistics as features, as well as performance metrics used by bloggers and statisticians employed by

teams. They used WEKA's implementations of ANN, Naïve Bayes, SVM and C4.5 for training classifiers on datasets describing National Hockey League (NHL) match results in the 2012/2013 season. The entire dataset amounted to 517 games. The authors concluded that traditional statistics outperformed the newer performance metrics in predicting the results of single games using 10-fold cross validation, while the ANN displayed the best accuracy (59%). Research into extracting more informative features and predicting the winners of NHL playoffs was cited as future work, as well as incorporating knowledge from similar sports such as soccer.

Weissbock and Inkpen (2014) combined statistical features with features derived from pre-game reports to determine whether the sentiment of these reports was useful for predicting NHL Ice Hockey matches. Data from 708 NHL matches in the 2012/2013 season were collected, and the pre-game reports were obtained from NHL.com. Both natural language processing and sentiment analysis based features were used in the experiments. The three statistical features, identified from their previous research (Weissbock et al., 2013), were: cumulative goals against and their differential, and the match location (home/away). The following algorithms from the WEKA toolkit were applied with their default parameters: ANN, Naïve Bayes, Complement Naïve Bayes, Multinomial Naïve Bayes, LibSVM, SMO, J48 Decision Tree, JRip, Logistic Regression, and Simple Logistic Regression. Three models were compared: models that used only the statistical features, models that used only the pre-game report text, and models trained with the features from sentiment analysis. It was found that models using only the features from pre-game reports did not perform as well as models trained with only the statistical features. A meta-classifier with majority voting was implemented, where the confidence and predicted output from the initial classifiers was fed into a second layer. This architecture provided the best accuracy (60.25%). This cascading ensemble approach on the statistical feature set provided superior performance to using both feature sets, suggesting that the pre-game reports and statistical features provided somewhat different perspectives. The authors commented that it was difficult to predict matches with a model trained using data from one or two seasons prior, probably due to player and coaching changes, etc.

Gu et al. (2019) reported that ensemble methods provided encouraging results in the prediction of outcomes of NHL Hockey matches over multiple seasons. The data extraction was automated and scraped from several websites containing NHL matches from the 2007/2008 to 2016/2017 seasons (1,230 matches). Data was merged from several sources including historical results, opposition information, player-level performance indicators, and player ranks using PCA. A total of 26 team performance variables were also included in their model. The kNN, SVM, Naïve Bayes, Discriminant Analysis and Decision Tree algorithms were applied, along with several ensemble-based methods. The ensemble-based methods (Boosting, Bagging, AdaBoost, RobustBoost) achieved the highest accuracy on the test set (91.8%). In terms of future research, the authors mentioned that additional data could further improve predictions, and different or additional features could be used, e.g., player ranking metrics, moving-averaged or exponentially-smoothed team and player performance metrics, psychological factors, the strategic or tactical judgements of coaches/experts, and players' physical or mental assessments. They also mentioned that the ML problem could be re-formulated so that a different outcome is predicted, e.g., whether a team will make

the playoffs or which team will win the championship, by training a model on the regular season data.

Paper	Competition	Models used	No. of features	No. of Matches	Accuracy of best model
Weissbock et al. (2013)	NHL	Naïve Bayes, SVM, ANN*, C4.5 Decision Tree	11	517	59%
Weissbock and Inkpen (2014)	NHL	Weka: ANN, Naïve Bayes, Complement Naïve Bayes, Multinomial Naïve Bayes, LibSVM, SMO, J48, JRip, Logistic, Simple Logistic, Simple Naïve Bayes, Cascading Ensemble*	6	708	60.3%
Gu et al. (2019)	NHL	KNN, SVM, Naïve Bayes, Discriminant Analysis, Decision Trees + ensembles of these: Boosting*, Bagging, AdaBoost, RobustBoost	19	1,230	91.8%

Table 6: Ice Hockey Studies (*denotes the best performing model).

3.1.6 Multiple Invasion Sports

McCabe and Trevathan (2008) used data from 2002 to 2007 and considered four different sports: Australian National Football League (NFL) Rugby League, Australian Football League (AFL) Australian Rules Football, Super Rugby (Rugby Union), and EPL soccer. ANNs trained with BP and Conjugative-Gradient Descent (CGD) were applied, with the former found to be slightly more accurate but had longer training time. The features used were the same across all of the four sports, i.e., sport-specific features derived from in-play events within matches were not included. The average accuracy achieved with the BP-trained ANN was 67.5%, higher than the expert predictions, which ranged from 60% to 65%. For future work, the authors mentioned that other sports could be considered, more features could be included, and the points margin could instead be predicted.

Reed and O'Donoghue (2005) predicted the results of EPL Soccer and English Premiership Rugby, building seven models including Multiple Linear Regression, ANN, and Discriminant Analysis. Their dataset consisted of the matches of three Soccer teams and two Rugby teams across three seasons, and contained seven features: match venue, rest, the positions of the team and opposition team in the league table, distances travelled to the match, and form. These features were used to predict both Rugby and Soccer matches, i.e., variables specific to in-play match events in Rugby or Soccer were not included. Given that a draw is a much more common result in Soccer than in Rugby, it is surprising that the accuracy obtained for Soccer (57.9%) was higher than for Rugby (46.1%). The models outperformed the predictions of human experts, and the ANN was found to achieve the best accuracy. The authors stated that motivational, injury and other variables could be included in future studies, and complex pattern recognition technology could be trialed rather than inflexible, simple linear models.

3.2 Striking & Fielding Sports

As opposed to invasion sports, which are time dependent, striking and fielding games are innings dependent. Perhaps the most widely-played striking and fielding sports are baseball, which is widely played in countries such as the United States, Japan, Korea and the Dominican Republic, and cricket, which is widely played in England and the countries of

Paper	Sport-Competition	Models used	No. of features	No. of Matches	Accuracy of best model
McCabe and Trevathan (2008)	Rugby League (NRL)	ANN trained with BP* and CGD	19	unknown	63.2%
McCabe and Trevathan (2008)	Australian Rules Football (AFL)	ANN trained with BP* and CGD	19	unknown	65.1%

Table 7: Other invasion sport studies (*denotes the best performing model)

the current/former British Commonwealth (e.g., Australia, New Zealand, South Africa, and nations in the Caribbean and Indian subcontinent).

3.2.1 Baseball

Valero (2016) performed a comparative study for prediction of Baseball matches, using 10 years of Major League Baseball (MLB) data. Lazy Learners, ANNs, SVMs and Decision Trees were the candidate models. The CRISP-DM framework was used as the experimental approach, and a classification approach (win/loss for the home team) was compared with a numeric prediction approach that predicted the run difference between the home and away teams. Feature selection and ranking methods in WEKA were applied to rank the original set of 60 features. The model used only the top three ranked variables: home field advantage, Log5 ratings, and the Pythagorean Expectation, and it was found that adding additional features did not improve results. The Pythagorean Expectation, developed by the Baseball statistician Bill James (James, 1984), represents the expected number of wins for a team given their runs scored and runs allowed, and Log5 ratings are essentially the same as Elo Ratings (Elo, 1978). SVM produced the best accuracy for both the classification and numeric prediction approaches. Consistent with the results of Delen et al. (2012), the classification approach performed significantly better than the numeric prediction approach. The SVM classification model achieved accuracy of around 59%. However, when using the 2005-2013 seasons as training data and the 2014 season as test data, the model's predictions were not significantly more accurate than predictions derived from match betting odds. The authors highlighted the difficulty in predicting outcomes in Baseball using statistical features alone, but suggested that experiments using the Japanese or Korean Baseball leagues could be useful. In future work, the authors also considered adjusting their model parameters, refining features, extending their datasets, and applying their model to other sports such as Basketball, Football, and Water Polo.

3.2.2 Cricket

Pathak and Wadhwa (2016) predicted the results of One-Day International (ODI) Cricket, and included four features based on the prior work of Bandulasiri (2008): toss outcome, match venue (home or away), time (day or night), and whether the team batted first or second. Three classification models were applied: Naïve Bayes, Random Forest and SVM. Data for matches from 2001 to 2015 were collected from cricinfo.com. A separate model was constructed for each team, analyzing that particular team with respect to all other teams. A training-test split of 80%:20% was used. To mitigate the effect of imbalanced datasets

Paper	Sport-Competition	Models used	No. of features	No. of Matches	Accuracy of best model
Pathak and Wadhwa (2016)	Cricket (ODIs)	Naïve Bayes, Random Forest, SVM*	4	unknown	61.7% (balanced accuracy)
Valero (2016)	Baseball (MLB)	Lazy Learners, ANN, SVM*, Decision Tree	3	24,300	59%

Table 8: Striking/Fielding sport studies (*denotes the best performing model)

(some teams had a high ratio of wins to losses), the three models were evaluated based on balanced accuracy and the Kappa statistic. SVM was found to perform the best across all teams, with an average balanced accuracy of 61.7%. It was suggested that, in future work, newer classification methods could be used, additional features could be included, and the approach could be applied to other forms of cricket, e.g., test matches and T20, and to other sports, e.g., baseball and football.

4. Discussion

In this section, we provide some discussion, from a machine learning perspective, on what can be observed across the surveyed articles. In Section 4.1, we discuss the most commonly applied ML techniques in this domain. In Section 4.2, we discuss data-related aspects of ML for team sport match result prediction, including feature selection and engineering, e.g., methods, dataset size, and feature subset comparison. We also highlight the fact that in-play features are specific to each sport, and that, in this respect, there are opportunities for inter-disciplinary collaboration between machine learning and sport performance analysis researchers. In addition, we discuss dataset size in terms of the number of instances, model training and validation approaches, cross-validation with chronologically-ordered instances, and class variable definition and comparison. In Section 4.3, we then describe evaluation of performance in terms of the most common evaluation metric (accuracy), as well as benchmark datasets and other ways to evaluate performance. Then, in Section 4.4, we discuss between-sport differences in predictability, based on the inherent characteristics and points-scoring systems of different sports. Finally, in Section 4.5, the most common future research directions that were extracted from the surveyed studies are analyzed and discussed.

4.1 Machine Learning Algorithms

ANNs are one of the most predominant ML techniques used in the analysis of data in sports (Schumaker et al., 2010). Indeed, in the course of our review, we found that a number of studies, especially early ones, used an ANN as the only predictive model, and did not compare its performance with any other models. This may well have been a consequence of the availability of specific software, tools and algorithms at the time of a study was conducted, rather than being motivated by a belief that ANNs are inherently better for

predicting match results in team sports. In a 2019 article in MIT Technology Review[1], it was found that, in an analysis of 16,625 AI-related papers, there has been a notable shift away from knowledge-based methods (those that derive rules or logic) over the past two decades. In the first half of the 2000s, there was an increase in use of ANNs, but from 2004 to 2014 their use declined. In the second half of the 2010s, ANNs again gained in popularity, probably due to the emergence of deep learning. Nonetheless, the common application of ANNs for team sport results prediction, often as the sole predictive model, prompts us to investigate whether the evidence in the literature suggests that ANNs have performed better than other models in practice.

Our research found that the majority of studies (65%) considered ANNs in their experiments, as shown in Figure 3. In earlier studies in particular, 23% of the papers considered ANNs as their only predictive model. The greater propensity for researchers to use ANNs in this domain has also resulted in the majority of studies attributing their highest accuracy to ANNs (Figure 5). It should also be noted that studies on two sports in this graph (Rugby League and Australian Rules Football) considered ANNs only. Three sports (American Football, Ice Hockey and Basketball), which found their best accuracy performances with alternative algorithms, also used ANNs, however, the alternative algorithms outperformed them. Therefore, the evidence does not suggest that ANNs have consistently performed better than other ML algorithms in predicting the match results of team sports. Indeed, the broader ML literature, real-world applications, and ML competitions (e.g., those held on Kaggle.com) do not support blanket statements that would give primacy to ANNs over all other algorithms. It is unclear why, historically, researchers displayed a preference for ANNs given that they are not straightforward to parameterize optimally and often over-fit, especially in the absence of sufficiently large datasets, which tends to be the norm rather than the exception in this domain. Another disadvantage of ANNs is that they are not easy to interpret and are therefore less useful to performance analysts and coaches seeking to draw out insights than other, more interpretable, ML models.

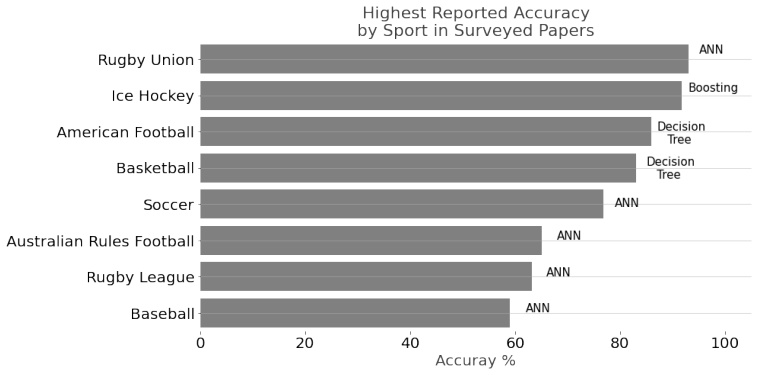


Figure 5: Highest recorded accuracies in studies by the different sporting codes covering the review period. The highest accuracy is reported along with the associated algorithm

1. https://www.technologyreview.com/2019/01/25/1436/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/

Decision Trees were the second most commonly applied technique (Figure 3). The appeal of these algorithms is obvious in that they are fast to train and usually do not possess a cumbersome number of tunable parameters. Importantly, they do not generate black-box models, but instead embody varying degrees of interpretability, depending on their implementation. This property can offer utility to professionals beyond just the ability to make predictions, but also in providing insight to coaches, management, and athletes. For instance, interpretable models like Decision Trees can aid in the identification of the most important performance indicator variables that influence match results, which is valuable in terms of developing appropriate strategies and focus areas. Their widespread use has resulted in both American Football and Basketball reporting their highest accuracy results using CART and Logistic Model Trees (LMT), respectively. Other variants of Decision Tree that have been popular in the literature include C4.5 (Quinlan, 1993) and its corresponding implementation in WEKA (Witten et al., 1999), J48.

Ensemble methods, Boosting algorithms, and Random Forests together form a top-three category of techniques used in the surveyed studies (Figure 3). Given the differing degrees of resilience of this family of algorithms to over-fitting, it is unsurprising that they have been used liberally, and have registered top accuracy results in Ice Hockey, in a recent study that highlighted the potential of ensemble-based solutions (Gu et al., 2019). A useful aspect of some of these models is their ability to achieve high accuracy while retaining interpretability. In future research in this domain, we see the potential for the application of Alternating Decision Trees (Freund & Mason, 1999), which possess the accuracy-improving benefits of boosting while retaining the interpretability of a decision tree structure.

Bayesian algorithms, e.g., Naïve Bayes and BNs, were one of the most popular sets of techniques used in the surveyed studies. Though these algorithms are not listed as performing the best in individual sports (Figure 5), they have been found in some comparative studies to offer better accuracies than alternative algorithms (Joseph et al., 2006; Miljković et al., 2010; Odachowski & Grekow, 2012). The popularity of Naïve Bayes in particular can be attributed to its common usage as a benchmarking algorithm when assessing the learnability of a new problem, which is tied to its ability to generate classifiers that do not over-fit. BNs were also shown to be useful when incorporating expert knowledge and when using small datasets (Joseph et al., 2006).

4.2 Data

In this subsection, the data-related aspects of the surveyed studies are discussed. In particular, we discuss feature selection and engineering, dataset size in terms of the number of matches/instances, model training and validation, and the use of cross-validation where instances are temporally ordered, which is the case for sports matches. Appropriate ways of defining the class variable, and comparing different class variables, are also discussed.

4.2.1 Feature Selection & Engineering

In early studies in particular, model features were often selected manually by researchers based on their knowledge of the specific sport. More recently, data-driven feature selection methods, including various filter-based techniques, have become more commonplace. These have ranged from CFS selection (Hall, 1998), to algorithms such as ReliefF (Kira & Rendell,

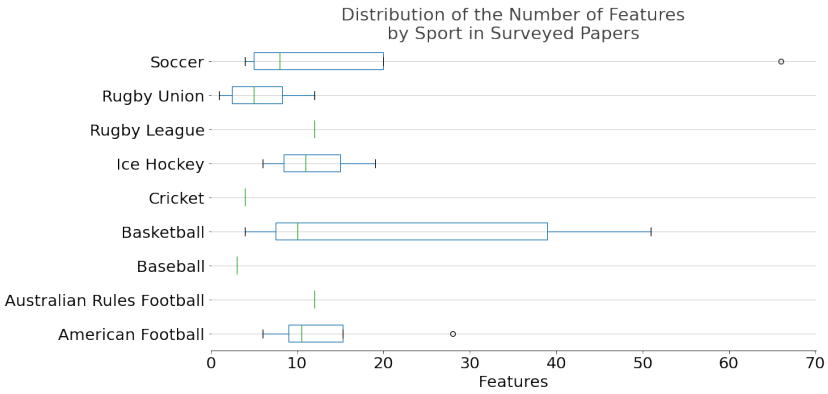


Figure 6: Distribution of the number of features used for machine learning by sport across all surveyed papers

1992), which have a greater contextual awareness and consider the existence of dependencies between features. Others have used feature-importance outputs from ML algorithms such as RIPPER (Cohen, 1995) to inform which features should be retained in the training of final models, while some (e.g., Loeffelholz et al., 2009) have used ANN-specific methods including signal-to-noise ratios (Bauer Jr et al., 2000). Sport experts have also been consulted to select what they consider to be the most important predictive features.

Compared to some other domains, the number of features used across the various sports have generally been on the modest side, and their distributions can be seen in Figure 6. If we introduce time as a dimension to discern if trends exist, we can see in Figure 7 that for Soccer, American Football, and Ice Hockey, a general tendency towards using larger feature sets can be observed. These are also sports that have, on average, seen some of the largest improvements in accuracies.

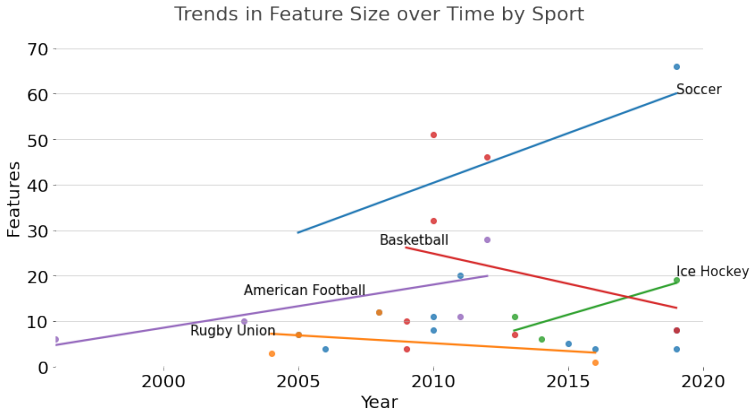


Figure 7: Differences in the distribution of predictive accuracies reported by sport.

In addition to the application of feature selection techniques, studies with a robust experimental process have generally compared several different feature subsets, e.g., betting odds, in-play features, features external to the match, player-level features, expert-selected

features, features extracted from pre-game reports, and features constructed from ratings. The performance of each feature subset can be compared, and can also be compared with the entire original set of features.

In-play features are sport-specific. Researchers have often proposed applying their predictive models to other sports. However, given that each sport has unique features that are associated with outcomes, it is generally not possible to directly apply a model to a dataset from a different sport. Rather, it is necessary to go through an entirely new experimental process on the dataset for that sport. Therefore, we recommend that, to approach prediction problems in a structured way, researchers should follow an experimental framework such as the CRISP-DM framework, the Knowledge Discovery in Databases (KDD) framework (Fayyad et al., 1996) or the Sport Result Prediction CRISP-DM (SRP-CRISM-DM) framework (Bunker & Thabtah, 2017) (the latter is an extension of CRISP-DM specifically designed for the application of ML to sport results prediction).

Feature selection & opportunities for inter-disciplinary collaboration. Nearly 90% of the studies we surveyed listed engineering additional richer features as one of their endeavours for future work (Figure 11). This should come as no great surprise, since generating more descriptive and therefore discriminatory features, with the help of domain knowledge, is generally accepted as the best strategy for improving predictive accuracy using ML (Domingos, 2012). The ability to generate more effective decision boundaries between instances of different classes is, to a larger degree, determined by richer features rather than by algorithms of increasing sophistication. To that end, domain expertise plays an important role in crafting more descriptive features. Such domain expertise could be obtained in consultation with coaches or athletes, or potentially from academic literature, e.g., from the field of sport performance analysis. Much of the research sport performance analysis considers so-called performance indicators (PIs). PIs represent the in-play features in sport result prediction models, and are usually augmented with features external to sport matches that may have an influence on outcome, e.g., weather, venue, travel, and player availability. Sport performance analysts, however, are not usually concerned with external variables because they are usually outside of the control of coaches and players. We suggest that there is an opportunity for knowledge transfer from the field of sport performance analysis, likely to be found in the identification and development of new model features. Given that feature engineering has been identified in this review as an area that is of highest priority for future research, greater collaboration between sport performance analysis and machine learning may result in meaningful advances. The two disciplines do not appear to have reached the point where a significant level of interchange is currently taking place, so researchers are encouraged to expand their collaborations in this respect.

4.2.2 Dataset Size

The distribution of the size of the datasets, in terms of number of instances, and their evolution over time, can be seen in Figure 8. It can be observed that, generally, the datasets have tended to be of a relatively small size, due to the limited amount of historical data in particular sports. This, together with fact that the data tends to be highly structured, potentially limits the ability of the datasets to capture the signal, which would likely result

in more accurate models. Sports that have seen an increase in dataset sizes are Basketball and, more acutely, Soccer.

It appears, however, that having access to a large dataset in terms of number of matches has not necessarily led to higher accuracy. This was particularly evident from the results of the 2017 Open International Soccer Database Competition, where model accuracies were modest despite the dataset containing over 216,000 matches. This particular pattern is depicted in Figure 9, where accuracy trends over time are rendered for all sports. Arguably, the growth in both the quantity and quality of features used in each sport (Figure 6) has grown more than the size of the datasets, which suggests that feature engineering and robust feature selection are likely to be key drivers of improvements in predictive accuracy.

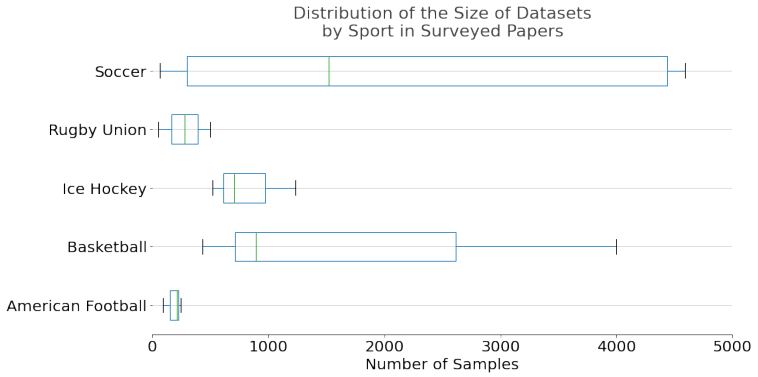


Figure 8: Distribution of the number of samples in datasets used for machine learning, by sport, across all surveyed papers.

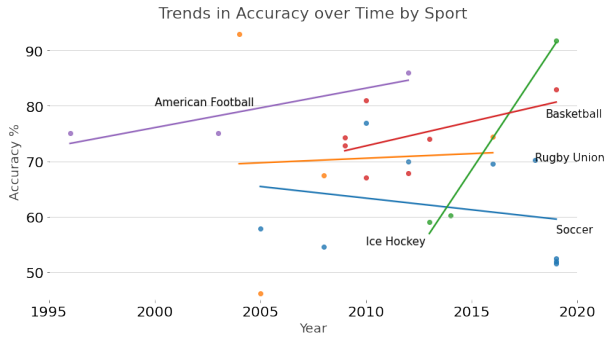


Figure 9: Trends in accuracy over time by sport.

4.2.3 Model Training & Validation

Successful studies have formulated experimental designs that have tested a number of different training-validation splits in their data. For instance, a certain number of historical seasons have commonly been used to train the model(s), and validation is performed on a current or future season. Or, if only one season of data was available, models are generally trained on a certain number of competition rounds and then validated on a future round in

that season. Researchers have often trialed various training-test splits and compared their accuracies.

4.2.4 Cross-Validation With Chronologically-Ordered Match Instances

Cross-validation was used in a number of studies. However, in match result prediction in sport, this can be problematic if the technique is not appropriately modified. Standard cross-validation randomly shuffles instances, so its application may mean that future matches are used to predict past matches. This pitfall was identified in a number of surveyed studies, and the predictive accuracies of these studies may have been compromised as a consequence.

4.2.5 Class Variable Definition & Comparison

A comparison of the performance of models, predicting both the points margin (home team points minus away team points) and discrete match outcomes (win/loss/draw or win/loss), has been investigated by a number of researchers (Delen et al., 2012; Valero, 2016; Danisik et al., 2018). Both approaches form a distinctive approach to formulating the ML task. However, results have been mixed, with no approach consistently shown to be superior. For instance, Delen et al. (2012) and Valero (2016) found that the classification approach performed better, while Danisik et al. (2018) found that a numeric prediction approach was superior on their dataset. Given these mixed results, we would recommend that, where possible, future researchers perform this type of comparison as part of their experimental process.

4.3 Performance Evaluation

In this subsection, common performance evaluation approaches, including evaluation metrics, are discussed. We discuss why predictive accuracy, the most common evaluation metric in this domain, is indeed an appropriate evaluation metric for team sport match result prediction. We also discuss benchmark datasets, and other ways in which model performance can be evaluated if no such benchmark dataset is available.

4.3.1 Evaluation Metric: Predictive Accuracy

Predictive accuracy is the performance evaluation metric that has been used by the vast majority of researchers in this domain. This is understandable since it is intuitive and interpretable, and datasets in this domain tend not to be overly imbalanced. In particular, when each instance in a dataset represents a match between two teams (home and away), there is generally a slight class imbalance in favor of the home team due to the well-known home advantage phenomenon. Over a whole season (or multiple seasons) in a competitive league, class imbalance generally will not exist to the degree that predictive accuracy becomes inappropriate to use as a performance evaluation metric.

4.3.2 Benchmark Datasets

A number of studies highlighted the inherent difficulties in comparing the results of studies in this domain. This difficulty arises since studies usually differ in at least one of the following dimensions: the sport(s) considered, the input dataset, the model predictors, the

class variable, and/or the matches, seasons or competitions considered. Part of the challenge lies in the scarcity of available benchmark datasets. Although this has been resolved for soccer, with the creation of the Open International Soccer Database (Dubitzky et al., 2019), as mentioned, this dataset is limited in that it does not include features derived from in-play events. Sport datasets are becoming increasingly available on websites such as Kaggle.com, and these could come to act as benchmark datasets for other sports in the future.

4.3.3 Other Ways to Evaluate Performance

Despite an absence of benchmark datasets in sports other than soccer, there are a number of other ways for researchers to evaluate their experimental results, e.g., by comparing their results to some baseline measure. Common approaches that have been used in the literature include comparing the predicted outcomes with:

- *Predictions derived from betting odds:* The outcome with the lowest betting odds is used as the class variable. Betting odds have also proven useful for match result prediction, even when included as the sole model predictor (Tax & Joustra, 2015). However, Hubáček et al. (2019a) pointed out that, if the purpose of the model is to generate profit through betting strategies, betting odds should not be included as a model predictor if one wants to "beat the house."

- *Predictions of experts:* ML model predictions can be compared to the predictions made by experts on the same set of matches, which are often published online or in the media.

- *ZeroR (Majority-class selection):* This is a simple classification rule that always selects the majority class. In most cases, this will predict a home-team victory due to the existence of the well-known home-advantage phenomenon.

- *Random prediction:* A randomly selected match outcome.

Given the specific nature of most datasets in sporting domains in terms of their time periods and features, comparisons with the above are useful for researchers when reporting their experimental results.

4.4 Between-Sport Differences in Predictability

Given the existing data and possible confounding factors, it is difficult to determine whether predicting match outcomes in certain sports is inherently more challenging than in others. Luck will always be a factor. Some research into disentangling the size of the luck (randomness) component from the skill component when predicting outcomes has already been conducted Aoki et al. (2017). The authors note that, in Soccer, the teams will win only 50% of the time when favoured by bettors, while the favored teams win 60% of the time in Baseball, and 70% of the time in both American Football and Basketball. Their own modelling research concludes that, out of four teams sports that they considered, Basketball appears to be the sport in which skill plays the largest role in the final results, and therefore has the most predictability. Basketball was followed in order by Volleyball, Soccer and Handball.

Possible hypotheses. Based on the data we have gathered, we offer some hypotheses with respect to the causes of the differences in predictability between the different sports. One possible explanatory factor may simply be that sufficiently large datasets and rich feature sets, which support high predictive accuracies, have not been equally available to researchers across different sports. Another possible reason is that sports have received highly imbalanced amounts of attention in the ML literature, which could in itself be due to the fact that non-ML techniques already perform adequately in predicting match results. For these reasons, it is inappropriate to attribute lower accuracies in sports that have received less research attention, e.g., Cricket, Rugby League, and Australian Rules Football (Figure 2), to something that is intrinsically more non-deterministic in these sports compared to other sports. However, a little more can arguably be inferred from sports such as Soccer, Basketball, American Football, Rugby Union and Ice Hockey, which have garnered more attention.

Low-scoring sports can have higher unpredictability. Soccer has received the largest share of research, yet its highest recorded predictive accuracy was 78% (Figure 5), and it comes fifth with respect to accuracy across all sports. Rugby Union, on the other hand, has received limited research attention, yet it ranks the highest of all surveyed sports, with an accuracy of 93% (O'Donoghue & Williams, 2004) recorded as its best outcome. Interpreting this perhaps requires some caution. Low-scoring sports tend to embody a higher degree of random chance as a determinant of results and this, in part, would explain some of the variation. This is also supported by Aoki et al. (2017), who showed that pure chance can be the single factor of outcomes in as many as 18% of matches in a season.

Lower competitiveness suggests higher predictability. The differential in accuracies highlighted above may, however, to some degree, be attributable to the characteristics of the sports and, even more, to the competitive contexts to which the predictive experiments were applied. For instance, Rugby Union is a much smaller global sport than Soccer, being played by fewer nations, while historically being dominated by a handful of them. Understandably, then, the best results reported for Rugby Union originated from a study of the 2003 Rugby World Cup, which suggests a context with a higher degree of predictability.[2] The context for Soccer and the competitions from which the results were collected were markedly different. In particular, the results were drawn from both national and international events in which the depth of competition was arguably greater, and that ultimately may have created conditions in which accurate prediction of results was less deterministic. Some evidence supporting the notion that a great deal of the predictability of sport results is naturally determined by the inherent depth of the competitions under observation, rather than on the sports themselves, is supported by Figure 10, which shows a much higher variability in the predictive accuracies for Rugby Union than for Soccer. We can see that in Rugby Union studies that considered national premiership (Reed & O'Donoghue, 2005) and international franchise competitions (McCabe & Trevathan, 2008), which exhibit a greater degree of equality between teams, the accuracy was correspondingly much lower, at 46% and 74%, respectively. In addition, Soccer is a low-scoring sport compared to Rugby Union

2. Rugby Union is, however, known to be becoming less predictable as the gap between low-ranked and high-ranked nations is narrowing over time and, consequently, more upsets are occurring. This was evident in the reduced predictability of the 2015 Rugby World Cup compared to previous tournaments (O'Donoghue et al., 2016).

and, returning to this point, we postulate that this may also be a contributing factor to generally lower accuracies in its studies. In low-scoring sports, there is a larger element of randomness in outcomes, which decreases the performance of predictive models.

Sports with more possible outcomes are less predictable. Associated with this is the higher likelihood of matches ending in draws. Given that draws are not improbable, many researchers predicting soccer match results have formulated the ML task as a three-class problem (win/loss/draw), rather than a binary-class (win/loss) problem. Of course, in general, as the number of classes increase, the learning problem becomes more difficult and thus accuracies tend to decrease.

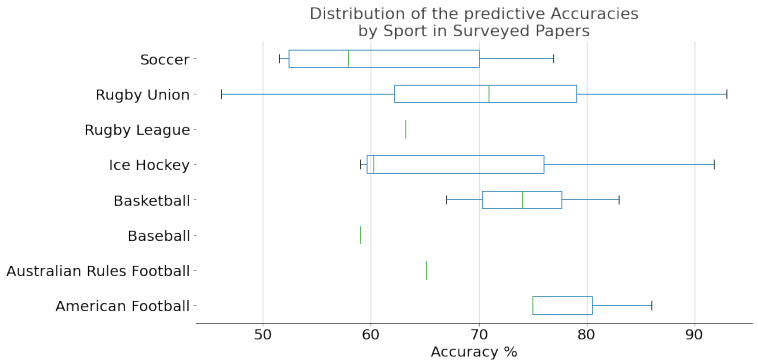


Figure 10: Differences in the distribution of predictive accuracies by sport.

Better features & richer datasets tend to increase predictability. Ice Hockey, on the other hand, is a counter-example, even though a binary-class approach was used in all studies. The sport is relatively low-scoring (although higher scoring than Soccer), yet its most recent and best result (92% accuracy) demonstrated considerably higher accuracy in comparison, as well as in comparison to American Football (86% accuracy) and Basketball (83% accuracy), which are both high-scoring sports. Studies on Ice Hockey, American Football and Basketball generally considered American national league competitions, e.g., the NHL, the NFL, and the NBA, which generally exhibit high degrees of competitiveness between teams. To some degree, this had the effect of controlling for the lopsidedness in expectations of outcomes that can exist in some sports and global competitions. When examining the trends over time in these three sports, some potential patterns emerged, hinting at an explanation for Ice Hockey's unexpectedly higher accuracies compared to the other two sports. We can see that the best performing result in Basketball had double the number of features of the previous study (74% accuracy Loeffelholz et al., 2009). However, the dataset was smaller and comparable ML algorithms were used, with standard Decision Trees performing the best out of the suite of methods explored. The best-performing result in American Football saw an 11 percentage-point improvement over the next best result in the NFL (Kahn, 2003), which may be attributed to a threefold increase in the total number of features used compared to the prior study, since both the total size of the dataset remained essentially the same, and the ML algorithms and training procedures did not represent a considerable increase in innovation. On the other hand, the leap from 60% accuracy (Weissbock & Inkpen, 2014) to 92% accuracy for Ice Hockey can be attributed to multiple factors. In particular, the authors used triple the number of features, double the

number of instances, and used more sophisticated algorithms and training approaches in the form of ensembling.

Effects of differing point-scoring systems on predictability. Across invasion sports, points-scoring systems and the manner in which points are attributed to scoring events differ, and this can affect the predictability of each sport. For instance, goals in Soccer increment the score by one, and it is not possible for a team to increment the score by more than one from any one scoring play. Ice Hockey is similar in this respect. On the other hand, it is possible in Basketball to increment the score by different amounts in each scoring play since there are three-pointers, in-circle shots (two points), and free-throws (one point). Different possibilities for increments in the score also exist in Rugby Union, Rugby League, American Football, and Australian Rules Football. Sports that have different possibilities for increments in score have more possible permutations in the final match scores and thus the result, an assertion that is also supported by Aoki et al. (2017).

Predictability is ultimately multi-factorial. In summary, although some invasion sports do embody characteristics such as being low-scoring (which makes outcomes harder to predict accurately, especially if a multi-class formulation is used), the competitive depth of the types of competitions in which the matches take place is likely to also be an important factor. A reasonable assumption is that, in general, the match outcomes of sports that are highly competitive, low-scoring, and have less possible increments in score will be more difficult to predict. However, some evidence does suggest that these difficulties can be mitigated to an extent when large and rich feature sets are used in conjunction with datasets containing a large number of instances, and cutting-edge training procedures are employed that combine multiple algorithms into robust, ensemble-based solutions.

4.5 Common Future Research Directions

Figure 11 depicts the percentage of surveyed studies that cited a specific future research direction (only research themes that were cited more than once are shown on the chart).

As mentioned, Figure 11 shows that early 90% of surveyed studies cited engineering additional features, which was by far the most common intended future research direction. Next, experimenting with alternative ML algorithms was cited as a future undertaking by nearly 40% of the surveyed studies. Each ML algorithm embodies within it assumptions about characteristics of the problem dataset, which may or may not hold, and the extent of the disconnect between the two will also affect the generalizability of different algorithms to varying degrees. A reasonable course of action, when an improvement in accuracy is sought, is to investigate different families of algorithms, which is the intention that Figure 11 appears to indicate that researchers are heading towards. What is more, non-parametric algorithms have a tendency to over-fit, and this is exacerbated on smaller datasets. A quarter of the studies signaled their intention to increase the size of their datasets, which, in this instance, would be the correct course of action for studies that have experienced this difficulty.

Meanwhile, improving training methods by reformulating the ML problem can also have a significant effect on accuracy, and nearly 40% of the studies intend to explore this. The types of training modifications that were cited for future work included using different combinations of previous season results (where appropriate) and giving greater weight to

datasets that are more recent. Some proposed applying clustering to the datasets before applying ML, and using alternative algorithms on different datasets.

Given that each sport has unique characteristics and potential inflection points that can act as markers for winning outcomes, it is somewhat surprising that a quarter of the studies plan to apply their ML methodology, including features, to other sporting codes. This seems rather counter-intuitive, given that 90% of these papers intend to generate more custom-designed features that are more tailored to their respective sporting codes.

Furthermore, it was also unanticipated to find that only 10% of the studies cited improvements in feature selection as an area of pursuit for future research. The problem of over-fitting is amplified when the pool of features is too large with respect to the size of the datasets. Given that most of the available datasets in this domain are not large, coupled with a strong consensus across all studies to pursue engineering and generation of more features, the problem of over-fitting is likely to plague many studies unless efforts to employ feature selection or dimensionality reduction techniques are afforded equal attention.

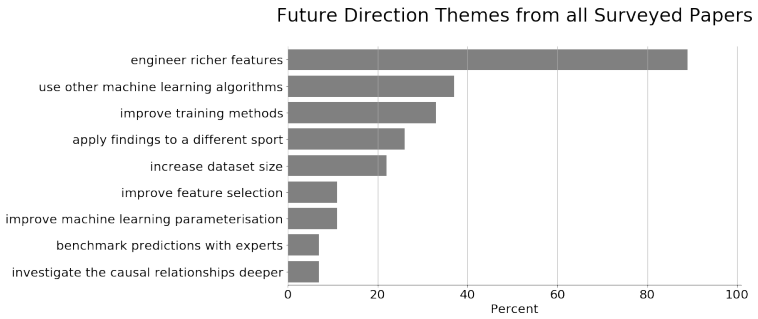


Figure 11: Commonly cited future research directions.

5. Conclusions

This survey reviewed studies published between 1996 and 2019 that applied machine learning (ML) methods to predict match results in team sport. In contrast to previous review articles in the same period, the scope of this review was defined to be narrow enough to allow for sufficiently in-depth analysis, but not overly wide so as to result in an unmanageable number of surveyed papers. The surveyed papers were categorized and sub-categorized by sport type (invasion sport or striking/fielding sport) and sport, respectively, and were analyzed and summarized in tabular formats. Various characteristics of the surveyed studies were analyzed including the types of models and experimental approaches used, the best performing models, the number of features included, and the total number of instances available. From this, we then discussed commonly applied ML models, data-related ML considerations including feature selection and engineering, dataset size, training/validation approaches, the use of cross-validation when considering the temporal order of match instances, and the definition of the class variable. Aspects related to performance evaluation from the surveyed papers were also discussed, including appropriate evaluation metrics, benchmark datasets, and other common evaluation approaches, e.g., predictions derived from betting odds, predictions of experts, majority-class selection, and random prediction.

We then discussed differences in the inherent predictability of sports due to their characteristics and points-scoring systems, before discussing the most common future research directions across the surveyed studies.

Overall, in this survey we found that, although ANNs were commonly applied especially in early studies, often as the sole predictive model, this may have been due to available software and tools at the time, since the evidence suggested (like other domains) that ANNs do not necessarily have primacy over other ML models. Thus, future researchers are recommended to compare a a set of candidate models.

Some recent studies (outside of the time-period considered in this review) have applied deep learning techniques to predict sport results (e.g., Chen et al., 2020 and Rudrapal et al., 2020); however, a limitation in this domain may be in their lack of interpretability. We suggest that Alternating Decision Trees (Freund & Mason, 1999), which combine the accuracy-boosting benefits of ensembles while retaining the interpretability of a decision tree structure, could be a suitable model to apply to predict match results in team sports.

Selecting an appropriate feature set and engineering additional features is crucial for prediction, and appears to be more important for accuracy than access to a large number of matches/instances. We see opportunity for greater collaboration between researchers from sport performance analysis and machine learning to define sport-specific in-play features, also known as performance indicators.

Researchers should consider their model training and validation approach, e.g., by using a hold-one-out approach to train their model on all prior matches to predict one future match, or by using historical seasons to predict the current season, or by using historical competition rounds to predict the current round (or a future round). Care needs to be taken if cross-validation is used since the standard method shuffles instances, which may result in future matches (inappropriately) being used to predict past matches.

Researchers also need to consider whether to use a discrete class variable (e.g., win/loss or win/loss/draw) or a numeric class variable (e.g., points margin). These two types of class variables were compared in a number of the surveyed studies, and this approach might warrant investigation by future researchers in this domain.

We found that accuracy is, by far, the most common evaluation metric used in this domain, which seems appropriate given that it is both interpretable and intuitive, and match result datasets generally are not particularly imbalanced. Comparing the accuracies of studies remains difficult even within the same sport due to different datasets, seasons, and predictive features being used. Such comparisons are even more difficult for studies that consider different sports with distinct characteristics and points-scoring systems.

Although a lack of benchmark datasets in this domain has been addressed to a certain extent for soccer through the Open International Soccer Database, this dataset is limited in that it does not contain in-play features. There remains no generally-accepted benchmark datasets for other sports. There are, however, other performance evaluation approaches with which researchers can gauge the performance of their models, e.g., by comparing their model predictions to those obtained from betting odds, from experts, or based on majority-class selection (usually a home team victory) or random prediction.

Due to the explosion in the number of papers published in the last decade in this domain, we recommend that future surveys focus on the application of machine learning for match result prediction in one specific sport and conduct a systematic review using,

e.g., the Preferred Reporting Items of Systematic reviews and Meta-Analyses (PRISMA) framework.

References

Aoki, R. Y., Assuncao, R. M., & Vaz de Melo, P. O. (2017). Luck is hard to beat: The difficulty of sports prediction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1367–1376.

Bandulasiri, A. (2008). Predicting the winner in one day international cricket. *Journal of Mathematical Sciences & Mathematics Education*, *3*(1), 6–17.

Bauer Jr, K. W., Alsing, S. G., & Greene, K. A. (2000). Feature screening using signal-to-noise ratios. *Neurocomputing*, *31*(1-4), 29–44.

Beal, R., Norman, T. J., & Ramchurn, S. D. (2019). Artificial intelligence for team sports: a survey. *The Knowledge Engineering Review*, *34*.

Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, *108*(1), 97–126.

Bunker, R. P., & Thabtah, F. (2017). A machine learning framework for sport result prediction. *Applied computing and informatics*.

Buursma, D. (2010). Predicting sports events from past results. In *14th Twente Student Conference on IT*.

Cao, C. (2012). Sports data mining technology used in basketball outcome prediction..

Carpenter, G., & Grossberg, S. (2003). Adaptive resonance theory, the handbook of brain theory and neural networks. *MA Arbib (ed)*, 87–90.

Chen, M.-Y., Chen, T.-H., & Lin, S.-H. (2020). Using convolutional neural networks to forecast sporting event results. In *Deep Learning: Concepts and Architectures*, pp. 269–285. Springer.

Cohen, W. W. (1995). Fast effective rule induction. In *Machine learning proceedings 1995*, pp. 115–123. Elsevier.

Constantinou, A. C. (2019). Dolores: a model that predicts football match outcomes from all over the world. *Machine Learning*, *108*(1), 49–75.

Constantinou, A. C., Fenton, N. E., & Neil, M. (2012). pi-football: A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, *36*, 322–339.

Constantinou, A. C., & Fenton, N. E. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, *9*(1), 37–50.

Danisik, N., Lacko, P., & Farkas, M. (2018). Football match prediction using players attributes. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pp. 201–206. IEEE.

David, J. A., Pasteur, R. D., Ahmad, M. S., & Janning, M. C. (2011). Nfl prediction using committees of artificial neural networks. *Journal of Quantitative Analysis in Sports*, *7*(2).

Davoodi, E., & Khanteymoori, A. R. (2010). Horse racing prediction using artificial neural networks. *Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing*, *2010*, 155–160.

Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting ncaa bowl outcomes. *International Journal of Forecasting*, *28*(2), 543–552.

Domingos, P. M. (2012). A few useful things to know about machine learning.. *Commun. acm*, *55*(10), 78–87.

Dubitzky, W., Lopes, P., Davis, J., & Berrar, D. (2019). The open international soccer database for machine learning. *Machine Learning*, *108*(1), 9–28.

Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2002). Modeling and prediction of competitive performance in swimming upon neural networks. *European Journal of Sport Science*, *2*(2), 1–10.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., et al. (1996). *Advances in knowledge discovery and data mining*, Vol. 21. AAAI press Menlo Park.

Freund, Y., & Mason, L. (1999). The alternating decision tree learning algorithm. In *icml*, Vol. 99, pp. 124–133. Citeseer.

Gu, W., Foster, K., Shang, J., & Wei, L. (2019). A game-predicting expert system using big data and machine learning. *Expert Systems with Applications*, *130*, 293–305.

Haghighat, M., Rastegari, H., & Nourafza, N. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal*, *2*(5), 7–12.

Hall, M. A. (1998). Correlation-based feature subset selection for machine learning. *Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, *29*(2), 147–160.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Huang, K.-Y., & Chang, W.-L. (2010). A neural network method for prediction of 2006 world cup football game. In *The 2010 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE.

Hubáček, O., Šourek, G., & Železnỳ, F. (2019a). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, *35*(2), 783–796.

Hubáček, O., Šourek, G., & Železnỳ, F. (2019b). Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, *108*(1), 29–47.

Hucaljuk, J., & Rakipović, A. (2011). Predicting football scores using machine learning techniques. In *2011 Proceedings of the 34th International Convention MIPRO*, pp. 1623–1627. IEEE.

Hughes, M. D., & Bartlett, R. M. (2002). The use of performance indicators in performance analysis. *Journal of sports sciences*, *20*(10), 739–754.

Ishi, M. S., & Patil, J. B. (2021). A study on machine learning methods used for team formation and winner prediction in cricket. In Smys, S., Balas, V. E., Kamel, K. A., & Lafata, P. (Eds.), *Inventive Computation and Information Technologies*, pp. 143–156, Singapore. Springer Singapore.

Ivanković, Z., Racković, M., Markoski, B., Radosav, D., & Ivković, M. (2010). Analysis of basketball games using neural networks. In *2010 11th International Symposium on Computational Intelligence and Informatics (CINTI)*, pp. 251–256. IEEE.

James, B. (1984). *The Bill James Baseball Abstract, 1984*. Ballantine Books New York.

Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, *19*(7), 544–553.

Kahn, J. (2003). Neural network prediction of nfl football games. *World Wide Web electronic publication*, 9–15.

Keshtkar Langaroudi, M., & Yamaghani, M. (2019). Sports result prediction based on machine learning and computational intelligence approaches: A survey. *Journal of Advances in Computer Engineering and Technology*, *5*(1), 27–36.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine Learning Proceedings 1992*, pp. 249–256. Elsevier.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting nba games using neural networks. *Journal of Quantitative Analysis in Sports*, *5*(1).

Maszczyk, A., Gołaś, A., Pietraszewski, P., Roczniok, R., Zajac, A., & Stanula, A. (2014). Application of neural and regression models in sports results prediction. *Procedia-Social and Behavioral Sciences*, *117*, 482–487.

McCabe, A., & Trevathan, J. (2008). Artificial intelligence in sports prediction. In *Fifth International Conference on Information Technology: New Generations (itng 2008)*, pp. 1194–1197. IEEE.

Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010). The use of data mining for basketball matches outcomes prediction. In *IEEE 8th International Symposium on Intelligent Systems and Informatics*, pp. 309–312. IEEE.

Min, B., Kim, J., Choe, C., Eom, H., & McKay, R. B. (2008). A compound framework for sports results prediction: A football case study. *Knowledge-Based Systems*, *21*(7), 551–562.

Mitchell, S. A. (1996). Improving invasion game performance. *Journal of Physical Education, Recreation & Dance*, *67*(2), 30–33.

Odachowski, K., & Grekow, J. (2012). Using bookmaker odds to predict the final result of football matches. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 196–205. Springer.

O'Donoghue, P., & Williams, J. (2004). An evaluation of human and computer-based predictions of the 2003 rugby union world cup..

Oliver, D. (2002). Basketball on paper. brassey's..

O'Donoghue, P., Ball, D., Eustace, J., McFarlan, B., & Nisotaki, M. (2016). Predictive models of the 2015 rugby world cup: accuracy and application. *International Journal of Computer Science in Sport*, *15*(1), 37–58.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web.. Tech. rep., Stanford InfoLab.

Pathak, N., & Wadhwa, H. (2016). Applications of modern classification techniques to predict the outcome of odi cricket. *Procedia Computer Science*, *87*, 55–60.

Prasetio, D. (2016). Predicting football match results with logistic regression. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pp. 1–5. IEEE.

Purucker, M. C. (1996). Neural network quarterbacking. *IEEE Potentials*, *15*(3), 9–15.

Quinlan, J. R. (1993). C4. 5: programs for machine learning..

Rajšp, A., & Fister, I. (2020). A systematic literature review of intelligent data analysis methods for smart sport training. *Applied Sciences*, *10*(9).

Razali, N., Mustapha, A., Utama, S., & Din, R. (2018). A review on football match outcome prediction using bayesian networks. In *Journal of Physics: Conference Series*, Vol. 1020, p. 012004. IOP Publishing.

Read, B., & Edwards, P. (1992). Teaching children to play games. *Leeds: White Line Publishing*.

Reed, D., & O'Donoghue, P. (2005). Development and application of computer-based prediction methods. *International Journal of Performance Analysis in Sport*, *5*(3), 12–28.

Rotshtein, A. P., Posner, M., & Rakityanskaya, A. (2005). Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, *41*(4), 619–630.

Rudrapal, D., Boro, S., Srivastava, J., & Singh, S. (2020). A deep learning approach to predict football match result. In *Computational Intelligence in Data Mining*, pp. 93–99. Springer.

Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). *Sports data mining*, Vol. 26. Springer Science & Business Media.

Shi, Z., Moorthy, S., & Zimmermann, A. (2013). Predicting ncaab match outcomes using ml techniques-some results and lessons learned. In *ECML/PKDD 2013 Workshop on Machine Learning and Data Mining for Sports Analytics*.

Somboonphokkaphan, A., & Phimoltares, S. (2009). Tennis winner prediction based on time-series history with neural modeling. In *Proceedings of the International Multi-Conference of Engineers and Computer Scientists*, Vol. 1.

Tax, N., & Joustra, Y. (2015). Predicting the dutch football competition using public data: A machine learning approach. *Transactions on Knowledge and Data Engineering*, *10*(10), 1–13.

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). Nba game result prediction using feature analysis and machine learning. *Annals of Data Science*, *6*(1), 103–116.

Tilp, M., & Schrapf, N. (2015). Analysis of tactical defensive behavior in team handball by means of artificial neural networks. *IFAC-PapersOnLine*, *28*(1), 784–5.

Tsakonas, A., Dounias, G., Shtovba, S., & Vivdyuk, V. (2002). Soft computing-based result prediction of football games. In *The First International Conference on Inductive Modelling (ICIM'2002). Lviv, Ukraine*. Citeseer.

Valero, C. S. (2016). Predicting win-loss outcomes in mlb regular season games–a comparative study using data mining methods. *International Journal of Computer Science in Sport*, *15*(2), 91–112.

Van Eetvelde, H., Mendonca, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of experimental orthopaedics*, *8*(1), 1–15.

Weissbock, J., & Inkpen, D. (2014). Combining textual pre-game reports and statistical data for predicting success in the national hockey league. In *Canadian Conference on Artificial Intelligence*, pp. 251–262. Springer.

Weissbock, J., Viktor, H., & Inkpen, D. (2013). Use of performance metrics to forecast success in the national hockey league.. In *MLSA@ PKDD/ECML*, pp. 39–48.

Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pp. 29–39. Citeseer.

Wiseman, O. (2016). *Using Machine Learning to Predict the Winning Score of Professional Golf Events on the PGA Tour*. Ph.D. thesis, Dublin, National College of Ireland.

Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with java implementations..

Zdravevski, E., & Kulakov, A. (2009). System for prediction of the winner in a sports game. In *International Conference on ICT Innovations*, pp. 55–63. Springer.

CHAPTER 8: PUBLICATION 5 - “A COMPARATIVE EVALUATION OF RATINGS AND MACHINE LEARNING-BASED METHODS FOR TENNIS MATCH RESULT PREDICTION”

This study considered performance at the match level in a dual sport: tennis. It compared Elo rating-based methods, namely conventional Elo ratings and Weighted Elo ratings (Angelini, Cantila, & De Angelis, 2022), against machine learning techniques. The previously proposed SRP-CRISP-DM (Bunker & Thabtah, 2019) framework was implemented and demonstrated in practice as the guiding framework for this study.

The first full year of ATP match data (2006) in a dataset containing matches from 2005 to 2020 was set to be the initial training set, and one year of data were incrementally added to this training set to predict 14 test years (2007 to 2020). The average rank of features was calculated based on five different feature selection methods. A comparative evaluation of the performance of five candidate machine learning models against Elo and Weighted Elo ratings was then conducted.

Of the five ML models utilised in the study, Alternating Decision Trees (ADTrees) (Freund & Mason, 1999) and logistic regression achieved higher accuracies than Elo ratings-based methods. These two models provided similar accuracies to match outcome predictions derived from betting odds. ADTrees provided strong performance and an interpretable decision tree structure, allowing average betting odds difference threshold variation.

The performance and interpretable structure of ADTrees may be helpful in future research and other applications in sporting contexts where interpretability is essential.

Bunker, R., Yeung, C., Susnjak, T., Espie, C., & Fujii, K. (2023). A comparative evaluation of Elo ratings and machine learning-based methods for tennis match result prediction. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology, 238(4), pp.305-316 © IMechE 2023. https://doi.org/10.1177/17543371231212235.

A Comparative Evaluation of Elo Ratings- and Machine Learning-based Methods for Tennis Match Result Prediction

Rory Bunker[1], Calvin Yeung[1], Teo Susnjak[2], Chester Espie[3], and Keisuke Fujii[1,4,5]

[1]Nagoya University, Japan
[2]Massey University, New Zealand
[3]Stetson University, USA
[4]RIKEN Center for Advanced Intelligence Project, Japan
[5]PRESTO Japan Science and Technology Agency, Japan

Introduction

Tennis is among the most popular sports in the world, as demonstrated by its levels of participation and spectatorship.[1] According to the International Tennis Federation's 2019 Global Tennis Report,[2] 87 million people (1.17% of the world's population) play tennis worldwide and, although figures are difficult to find in academic literature, tennis consistently appears in the top ten most popular global sports in online lists in terms of both participation and spectatorship. Due to the sport's popularity, there is great interest in forecasting the results of matches, for example, from fans and bookmaking companies. Bettors are interested in match result prediction to decide whether to place bets on matches. Bookmakers use predictive models as well as betting volumes to ensure their match betting odds are appropriate. Coaches and sport performance analysts are interested in interpretable models that allow them to identify the most important features that have contributed to wins in the past, or that players can alter in the future to improve their chances of winning.

Approaches to forecast match results in tennis include ratings-, regression-, points-, and paired comparison-based methods,[3] as well as machine learning models. Ratings-based methods use the difference in player ratings to identify the probable winner of a match, with ratings updated over time based on actual match results. Machine learning (ML) models are trained on a specific number of historical matches and are then tested on an unseen test set to confirm their generalizability. ML models belong to two broad groups: classification models and regression models, which aim to predict discrete and numeric target variables, respectively. In classification, which is considered in the current study, a set of input features is used to predict a labelled binary outcome variable (for example, win or lose). In regression, the target variable might be the margin-of-victory in terms of sets won, that is, the number of sets won by one player minus the number of sets won by the opposition player in a particular match[4] (Games won can also be used as the margin-of-victory; however, a player can win a match but have won fewer games than the opposition).

Elo ratings, which were originally used for the analysis of chess players,[5] have been commonly used to estimate team ratings and predict outcomes in various sports including Football, Australian Rules Football, and Gaelic Football.[6–12] Elo has been extended, for example, to incorporate home advantage,[7] margin-of-victory,[4,6] form/momentum,[13] and different playing surfaces, for example, grass, hard and clay courts.[14] Kovalchik[3] compared the performance of several models in predicting 2,395 Association of Tennis Professionals (ATP) tennis matches in 2014 and found that Elo performed better than regression-, points- and paired comparison-based models. Weighted Elo (WElo), recently proposed by Angelini et al.[13], extends standard Elo to consider form by incorporating a player's most recent match result. WElo was applied to over 60,000 ATP and Women's Tennis Association (WTA) matches, and it was found to outperform paired comparisons, logistic and probit regression, and standard Elo.[13] However, whether Elo and/or WElo can outperform common ML models, for

example, those used in a recent study by Wilkens:[15] boosted trees, random forests (RFs), support vector machines (SVMs), and artificial neural networks (ANNs), is an open question. In this study, a comparative evaluation of Elo/WElo against ML models is conducted using the same ATP match data as Angelini et al.[13]

Alternating decision trees (ADTrees/ADTs)[16,17] have been applied in various domains such as natural language processing,[18] medicine,[19,20] and flooding/land-slide susceptibility modeling and assessment,[21–23] but have not yet been applied in the context of sports. ADTrees are used as the boosting method in this study.

This study has three main contributions. First, a comparative evaluation of Elo/WElo with ML models is conducted, which is useful for researchers and practitioners in sports match result forecasting. Second, ADTrees are introduced in sports match result forecasting, which, to our knowledge, have not previously been applied for this purpose. Third, the previously proposed SRP-CRISP-DM[24] framework is demonstrated in practice.

The remainder of this paper is organized as follows. In the next section, background is provided on studies that have used ML or ratings to predict tennis match results. Then, Elo and WElo are described, as is how the SRP-CRISP-DM framework will be applied. The results are then presented and discussed before concluding the paper.

Background

In this section, first, studies that have used ML models for tennis match result prediction are reviewed, followed by studies that have used paired comparison/rating methods.

Machine Learning for Tennis Match Result Prediction

Logistic regression was commonly applied in early studies. Clarke and Dyte[25] used data from the Men's 1998 Wimbledon, 1998 US Open, and 1999 Australian Open Grand Slam tournaments, and fitted a logistic regression model to official ATP rankings to estimate a player's chance of winning as a function of the rating points difference. Klaassen and Magnus[26] applied logistic regression at both match- and point-levels to forecast the match winner at the beginning of, and during, Wimbledon grand slam matches. The estimated win probabilities were functions of the player ranking difference. Lisi[27] used logistic regression to predict 501 tennis matches using ATP points/rankings, the players' ages, home advantage, and information derived from bookmaker odds. The model was estimated using tournament data from 2012 and was applied to Grand Slam tournament matches from 2013, and the model-implied odds were used to place out-of-sample bets. Makino, Odaka and Kuroiwa[28] calculated the frequency of single shots, two-shot patterns, and effective shot patterns from individual points in ATP matches. An L1-regularized logistic regression was then used to predict the winners of points and to obtain useful tactical features. The highest accuracy achieved was 66.5%, and the most relevant features were the respective ratios of shots played by each player.

Recently, ML models other than logistic regression have been applied for tennis match result prediction. Ghosh, Sahu and Biswas[29] compared the performance of a decision tree, learning vector quantization, and an SVM to predict tennis match results, using eight UCI databases of Grand Slam tournament results. The decision tree was found to perform best, achieving 99.14% and 98.45% accuracy for a 70%/30% training/test split and 10-fold cross-validation, respectively. The very high accuracy reported in this study relative to others may indicate that future matches were incorrectly used to predict past matches. Gu and Saaty[30] predicted tennis match results using data and judgments. An analytic network process[31] that incorporated factor analysis and clustering was applied to 63 men's and 31 women's 2015 US Open matches, and achieved 85.1% accuracy. Candila and Palazzo[32] aimed for accurate predictions and profitable bets on men's tennis matches, applying an ANN to a multitude of features. When predicting out-of-sample, the ANN outperformed four of the five other models, regardless of the time period considered. Gao and Kowalczyk[33] used a random forest model to predict tennis match results based on player and match characteristics. Using data from 14 Grand Slams, serve strength, followed by break point conversion rate and the difference in player ranking points, were found to be the most important predictors of match results. Wilkens[15] applied ML methods to tennis matches and found that average prediction accuracy tends to level off at about 70%. The author also found that, regardless of the model applied, most relevant information was contained in betting odds, and adding other match- or player-specific information did not improve results. The betting odds-implied benchmark and set of models (logistic regression, neural network, random forest,[34] boosting, and SVM[35]) used by Wilkens will be used in the current study to compare with Elo/WElo.

Ratings & Paired Comparison Methods for Tennis Match Result Prediction

In this subsection, studies are reviewed that used rating and paired comparison methods (which are related) to predict tennis match outcomes.

Williams, Liu and Dixon[14] proposed a surface-specific Elo that considers different playing surfaces such as grass, hard and clay courts. Kovalchik[3] compared the performance of several models in predicting 2,395 ATP tennis matches in 2014 and found that Elo performed better than regression-, points- and paired comparison-based models. While Elo ratings have traditionally been updated based on discrete wins/losses in tennis, Elo has been extended to incorporate the margin-of-victory (MoV) by Kovalchik.[4] The author compared four different approaches to incorporate MoV: linear, joint additive, multiplicative, and logistic regression. Although all four approaches were found to outperform standard Elo, the joint additive model was preferred because a simulation study showed it had the most stable variance and smallest bias in terms of player rating differences. Angelini, Candila, and De Angelis[13] noted that standard Elo does not fully take into account current player form and their recent performances. They proposed Weighted Elo (WElo), which places additional weight on the scoreline of the player's last match to account for the "hot-hand" phenomenon (also called "momentum"). With a dataset of over 60,000 ATP and WTA matches, regardless of the training and test periods considered, WElo was found to outperform four competing methods: standard Elo, the Bradley-Terry paired comparison model, the logit regression of Klaasen and Magnus,[26] and the probit regression of Del Corral and Prieto-Rodriguez.[36] The authors also demonstrated that the MoV method of Kovalchik[4] is, in fact, a special case of WElo.

Bradley-Terry[37] paired comparison models are related to Elo because the Elo rating can be expressed as a function of the player's Bradley-Terry strength (subsection 2.1, Coulom[38]). McHale and Morton[39] proposed a Bradley-Terry model that incorporates current and historical player performance measures, and surface-specific parameters. Baker and McHale[40] improved upon this model by proposing a time-varying Bradley-Terry model that improves the ranking of tennis players by allowing players' abilities to vary over time (with player ability following a non-parametric function of time). Another recent study to have used a Bradley-Terry-based approach to forecast tennis match results is that of Fayomi, Majeed, and Algani.[41] A weakness of paired comparison models such as Bradley Terry is that when the number of player pairs is large, sparsity becomes an issue, a problem that some recent studies[42,43] have attempted to address.

Materials & Methods

In this section, Elo and WElo are described in the first subsection, and the experimental framework used in this study, the SRP-CRISP-DM framework, is described, along with how it will be applied, in the following subsection.

Elo and WElo Ratings

Suppose two players, player i and player j, with Elo ratings $R_i(t)$ and $R_j(t)$, respectively, play a match at time t. In standard Elo, the probability that player i beats player j is estimated (in what is known as the Estimation- or E-step) using the logistic curve:

$$p_{i,j}(t) = \frac{1}{1 + 10^{\frac{-(R_i(t) - R_j(t))}{400}}}$$

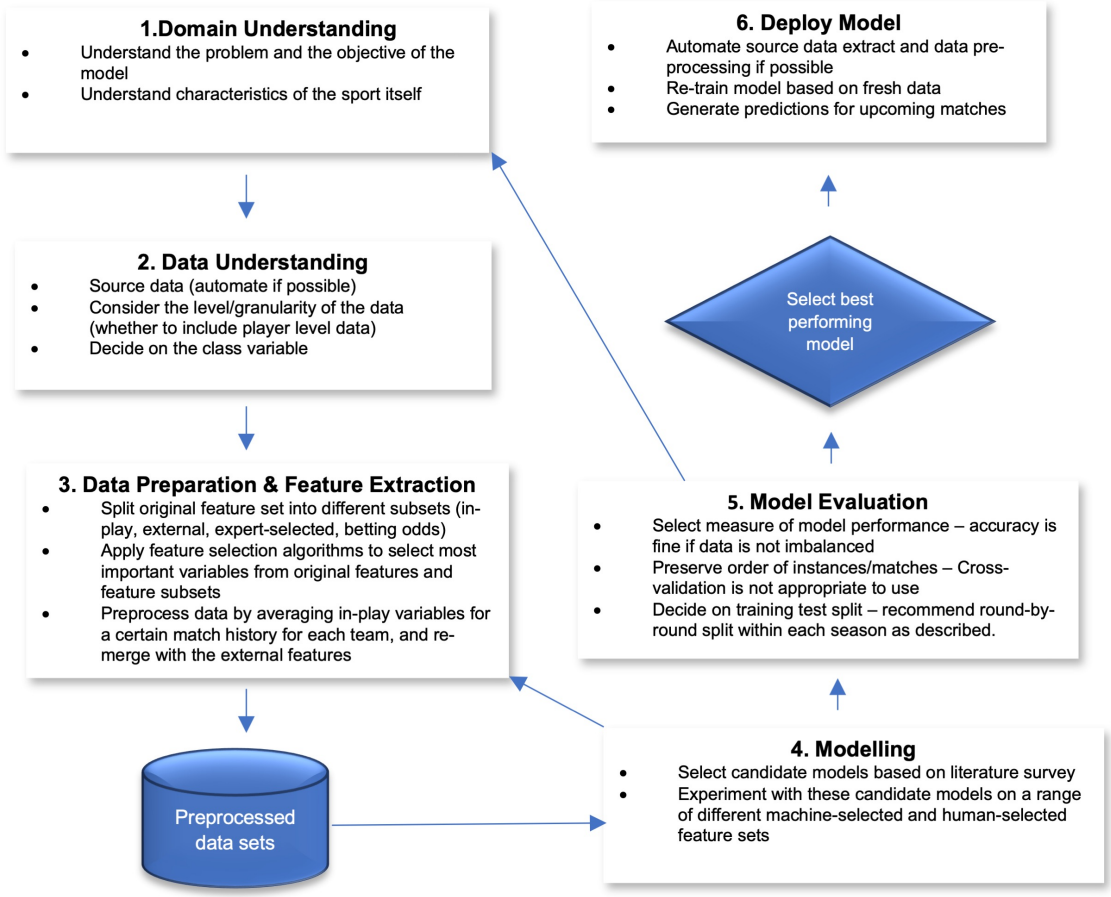
Similarly, the probability that player j beats player i is:

$$p_{j,i}(t) = \frac{1}{1 + 10^{\frac{-(R_j(t) - R_i(t))}{400}}}$$

Because these are probabilities, they add to one, that is, $p_{i,j}(t) + p_{j,i}(t) = 1$. A function representing the actual match result, $A_i(t)$, takes the value 0 if player i loses the match and 1 if player i wins the match, and player i's ratings are updated based on the following update step (U-step):

$$R_i(t+1) = R_i(t) + K_i(t)[A_i(t) - p_{i,j}(t)]$$

Figure 1. Steps of the Sports Result Prediction CRISP-DM (SRP-CRISP-DM) model, which extends the CRISP-DM framework for the specific problems faced in sports match result prediction tasks (two arrows that were not shown in the original paper[24] have been added to the figure).



In WElo,[13] player i's WElo rating, $E_i(t)$, is instead updated according to the following U-step:

$$E_i(t+1) = E_i(t) + K_i(t)[A_i(t) - p_{i,j}(t)]f(G_{i,j}(t))$$

where $f(G_{i,j}(t))$ is a function that depends on the number of games (or sets) won by each player in their previous match. Specifically, $f(G_{i,j}(t)) = \frac{N_i(t)}{N_i(t)+N_j(t)}$ if player i won match t and $f(G_{i,j}(t)) = \frac{N_j(t)}{N_i(t)+N_j(t)}$ if player i lost match t.

A fixed value of $K_i(t) = \bar{K} = 32$ has often been used in the literature. However, Kovalchik[3] found that an Elo rating system devised by FiveThirtyEight,[44] with values of $c = 250$, $o = 5$ and $s = 0.4$, in which K decreases with the number of matches a player has played in their career, performed best against several competing methods:

$$K_i(t) = \frac{c}{(M_i(t)+o)^s} = \frac{250}{(M_i(t)+5)^{0.4}} \quad (1)$$

where c is a constant, $M_i(t)$ is the number of matches in the dataset up to time t that have involved player i, o is a small offset that avoids very large values when $M_i(t)$ is small, and s is a curve shape-altering parameter. The "Kovalchik" K-value in Equation 1, with the above-mentioned FiveThirtyEight values, which is the default option in the WElo R package,[45] is used in the current study.

Experimental Framework: SRP-CRISP-DM

The SRP-CRISP-DM framework[24] extends the original CRISP-DM framework[46] for the specific issues faced when forecasting match results in sports. Like CRISP-DM, the SRP-CRISM-DM framework consists of six steps, which are illustrated in Figure 1.

In this study, the SRP-CRISP-DM framework steps are applied as follows.

1. Domain Understanding: In the current study, the objective of Elo/WElo and the ML models is to predict the results of tennis matches. Although one use case of match result prediction models is to devise profitable betting strategies through betting on sports matches, devising such strategies is not the focus of the current study.

2. Data Understanding:

Data Collection. Elo/WElo and the ML models were applied to the ATP men's dataset used by Angelini, Candila, and De Angelis,[13] which was originally sourced from tennis-data.co.uk and contains ATP matches from 5 July 2005 to 22 November 2020. The ATP data were downloaded as an RData file from the "Appendix C. Supplementary materials" section of the paper.[13] An R script was written that cleans the data (using the 'clean()' function in the WElo R package[45]) and then converts and exports it to a CSV file. The data cleaning reduced the number of matches from 38,868 to 33,976.

3. Data Preparation & Feature Extraction:

Data Preparation. Due to the structure of the data used for the Elo and WElo calculations in the initial data, the dataset needed to be transformed to apply the ML models. In particular, the initial dataset was structured such that each instance represents a specific match, with the winner always appearing in the column preceding the loser. The dataset with the winner always preceding the loser meant that a class/target variable with win/loss could not be created, which is necessary to apply supervised ML models. Therefore, a transformed dataset for ML was created so that the player order was arranged whereby, for each match, Player 1 "P1" and Player 2 "P2" columns were created, where P1 is the first player in terms of alphabetical order based on their surname. The target variable was defined as the match result from Player 2's perspective, that is, whether Player 2 won or lost against Player 1. Alphabetical ordering, based on surname, was designed to produce a dataset that was random and also balanced in terms of the target variable. Indeed, in the transformed dataset, the number of matches belonging to each of "P2 win" and "P2 loss" was split nearly 50:50, with 16,841 P2 wins (49.6%) and 17,135 P2 losses (50.4%). The features for ML were then created to be the differences between Player 1 and Player 2 (Player 1 minus Player 2).

Feature Extraction. Betting odds have proven to be useful for match result prediction in sports, even when used as the only feature in a model.[47] However, it has been pointed out that if the purpose of the model is to generate profit through betting strategies, betting odds should not be included as a model feature if one wants to "beat the house".[48] Because, as mentioned above, our purpose here is purely match result prediction, that is, profitable betting strategies are not attempting to be generated, it is considered appropriate to include betting odds as a model feature. As shown in Table 1, the candidate features were grouped into four types: in-play features (derived from events that occurred within matches), external features, and date and target (class) features. There are some potentially redundant features in the dataset. For example, the difference in Bet365 odds (B365OddsDiff) is correlated with AvgOddsDiff, the difference in average odds across 11 betting companies including Bet365 (Stoke-on-Trent, United Kingdom). In other words, Bet365 is one of the betting companies over which the average odds are calculated. Several ranking and ATP points features were also engineered: the absolute difference in the ATP points and rankings, the percentage difference in the ATP points and rankings, and the difference in ATP points and rankings normalized to be between -1 and 1. It was subsequently realized that the in-play features could not be used for match result forecasting, because the values of these features are, of course, not known until after matches have been played (unless some form of historical averaging process was applied to take the average of each of these in-play features across a certain number of historical matches, which, for simplicity, was not undertaken in this study). Nonetheless, how the values of the six in-play features differ between Player 2 wins and losses is of interest from a practical perspective and is presented in the supplemental material (Table 7). The remaining (external) features were used to predict the target variable because the date feature was only used for splitting the dataset into training and test sets.

Feature Selection. Five different feature selection methods from WEKA (correlation, chi-squared, information gain ratio, reliefF,[49] and symmetrical uncertainty attribute evaluation) were applied to the remaining features and their average rank was taken (Table 2). As mentioned, because the average betting odds difference (AvgOddsDiff) was calculated using the average odds taken across 11 companies including Bet365, it is correlated with the Bet365 betting odds difference (B365OddsDiff). As a result, only AvgOddsDiff, which had a higher average rank than B365OddsDiff, was ultimately used in the final ML models. Likewise, only one of the features derived from ATP rankings or ATP points (PtsDiff, RankDiff, PtsDiffPerc, RankDiffPerc, PtsDiffNorm, and RankDiffNorm) was used in the final ML models. In particular, because PtsDiffNorm, the ATP points difference normalized to be between -1 and 1, had the highest average rank among these six features, it was

selected for the final ML models. The final set of features used in the ML models was: AvgOddsDiff, P1 Seed, P1 Entry, SameHand, HeightDiff, SameCountry, AgeDiff, and PtsDiffNorm.

Table 1. Features in the initial transformed dataset, categorized into in-play and external features. *The date variable was used only for creating the training-test splits.

Feature	Description	Type
AcePercDiff	Difference in the percentage of points in the match that were aces	In play
DFPercDiff	Difference in the percentage of points in the match that were double faults	In play
FirstServePercDiff	Difference in first serve success percentage	In play
FirstServePtWinPerc	Difference in the percentage of points won when the first serve was successful	In play
PtsWonOnScndServePerServiceGameDiff	Difference in the points won on second serve per service game	In play
BreakPtSavePerc	Difference in the percentage of break-points saved	In play
Date*	Date of the match (used for the training-test split but not as a model feature)	Date
RankDiff	Difference in players' ATP ranks	External
PtsDiff	Difference in players' ATP competition points	External
RankDiffPerc	Percentage difference in players' ATP ranks	External
PtsDiffPerc	Percentage difference in players' ATP competition points	External
RankDiffNorm	Difference in players' ATP ranks, normalized to be between -1 and 1	External
PtsDiffNorm	Difference in players' ATP competition points, normalized to be between -1 and 1	External
B365OddsDiff	Difference in Bet365 odds	External
AvgOddsDiff	Difference in the average odds across 11 betting companies including Bet365 (Bet365, Bet&Win, Centrebet, Expekt, Ladbrokes, Gamebookers, Interwetten, Pinnacles Sports, Sportingbet, Stan James, and Unibet)	External
P1_Seed	Whether P1 was seeded for the tournament, and if so, what seed they were	External
P1_Entry	Whether P1 had special entry to the tournament, for example, whether they were a qualifier or wild card	External
SameHand	Whether P1 plays with the same hand as P2	External
HeightDiff	Difference in players' heights	External
SameCountry	Whether the players are from the same country	External
AgeDiff	Difference in players' ages (at the time the match was played)	External
P2_Result	Match result from player 2's perspective (that is, "W" if P2 beat P1, "L" if P2 lost to P1)	Target

4. Modelling:

The ML models used in the recent study of Wilkens[15] were compared with Elo/WElo: namely, logistic regression, ANN, random forest, boosting, and SVM. For the boosting method, ADTrees were employed. All models were applied in WEKA 3.9.6 with their default parameters.

Table 2. Feature rank in terms of correlation, chi-squared, information gain ratio, symmetrical uncertainty, and ReliefF (all available in WEKA 3.9.6), along with the average rank across these different methods.

Feature	CorrRank	ChiSqRank	GainRatioRank	SymmURank	ReliefFRank	AvgRank
AvgOddsDiff	1	2	1	1	1	1.2
B365OddsDiff	2	1	2	2	4	2.2
PtsDiffNorm	3	5	5	5	2	4
PtsDiff	4	6	6	6	3	5
RankDiffPerc	8	3	3	3	9	5.2
PtsDiffPerc	7	4	4	4	10	5.8
RankDiffNorm	5	7	7	7	8	6.8
RankDiff	6	8	8	8	7	7.4
P1_Seed	9	9	9	9	11	9.4
HeightDiff	11	11	11	11	6	10
P1_Entry	10	10	10	10	14	10.8
SameHand	12	12	12	12	12	12
AgeDiff	13	14	14	14	5	12
SameCountry	14	13	13	13	13	13.2

The ML models were compared with both Elo and WElo, which are available in the WElo R package.[45] The R script used to apply Elo and WElo is available on GitHub (see the supplemental material).

5. Model Evaluation:

Training-Test Splits. Match result prediction models are often built by training them on a certain number of historical matches and then predicting current or future matches. Because the first year of ATP data (2005) did not include a full year of data, it was excluded from the analysis. The year 2006 was chosen as the initial training set, and one year was incrementally added to this training set and used to predict the subsequent year. That is, the ML models were trained on 2006 and tested on 2007, trained on 2006-2007 and tested on 2008, trained on 2006-2008 and tested on 2009, and so on. Finally, the ML models were trained on 2006-2019 and tested on 2020. Therefore, there were a total of 14 training-test splits. Standard cross-validation randomly shuffles instances; therefore, it cannot be used in match result prediction because future matches will erroneously be used to predict past matches. An option in WEKA was specified to ensure that the temporal order of instances was maintained to ensure only historical matches are used to predict future matches (specifically, the "Preserve order for % Split" option was checked in the "Classifier Evaluation options" in the WEKA workbench under "test options" then "More options...").

Model Evaluation Metric. The imbalance ratio in the dataset was 0.983, which indicates that the size of the minority class was very similar to the size of the majority class. Because the dataset was not unbalanced, accuracy was deemed an adequate measure of model performance.

6. Deployment:

In this study, the models are essentially static in the sense that they do not need to be deployed in any form of production environment and re-run on fresh data.

Results & Discussion

The results of the accuracies of the ML models and Elo/WElo for the 14 training-test splits are shown in Table 3. Also shown in the table are the test set predictions derived purely from the betting odds, that is, based on the rule:

```
if AvgOddsDiff < 0 then P2 result = "L";
else P2 result = "W"
```

It should be noted that, of the 2,350 instances in 2010, 824 (35%) had null values for the AvgOddsDiff feature, and this high proportion of null values in the AvgOddsDiff feature reduced the accuracies achieved by some models (ADTree and LR) in that year. The results show that ADTree achieved the highest accuracy in seven of the 14 test sets. Predictions derived from betting odds achieved the best accuracy in six of the 14 test datasets. Logistic regression achieved the highest accuracy in one of the test sets (2016). The betting odds and ADTree achieved the equal-highest accuracy performance (72.7%) across all test sets. The results obtained are broadly consistent with past studies, for example, that by Wilkens,[15] who found that regardless

Table 3. Accuracy results (%) of the ML models and Elo/WElo.

Train	Test	#Train	#Test	Train%	Test%	LR	ANN	SVM	RF	ADT	Elo	WElo	Odds
2006	2007	2,306	2,368	49%	51%	70.9	65.1	63.7	68.9	71.0	67.3	66.4	**71.5**
2006-2007	2008	4,674	2,242	68%	32%	70.2	65.2	63.9	66.1	**70.9**	67.4	66.7	70.8
2006-2008	2009	6,916	2,278	75%	25%	70.3	65.8	66.5	67.8	**71.2**	69.7	69.6	70.8
2006-2009	2010	9,194	2,350	80%	20%	62.2	63.5	64.6	67.4	63.8	68.0	68.6	**70.8**
2006-2010	2011	11,544	2,345	83%	17%	71.6	69.3	68.6	66.6	71.4	69.3	69.3	**72.3**
2006-2011	2012	13,889	2,324	86%	14%	72.1	69.7	68.1	68.9	72.6	71.2	70.7	**72.7**
2006-2012	2013	16,213	2,316	88%	12%	69.2	67.8	67.6	66.5	68.9	67.3	68.2	**69.4**
2006-2013	2014	18,529	2,250	89%	11%	71.1	68.6	69.6	66.9	70.9	68.9	68.7	**71.3**
2006-2014	2015	20,779	2,285	90%	10%	71.7	66.1	71.4	70.4	**72.7**	69.4	69.3	72.5
2006-2015	2016	23,064	2,297	91%	9%	**71.4**	70.5	70.3	67.7	70.3	68.6	69.0	71.0
2006-2016	2017	25,361	2,307	92%	8%	67.8	65.8	66.7	66.4	**68.0**	67.0	66.5	67.9
2006-2017	2018	27,668	2,008	93%	7%	67.7	66.9	66.1	63.5	**67.9**	63.4	64.0	67.6
2006-2018	2019	29,676	2,303	93%	7%	66.8	66.2	65.5	64.6	**67.3**	63.7	64.0	66.9
2006-2019	2020	31,979	1,033	97%	3%	69.4	67.5	69.6	64.9	**69.8**	63.6	64.1	69.7
						69.5	67.0	67.3	66.9	69.8	67.5	67.5	**70.4**

Note: The highest accuracy for each training-test split is shown in bold, as is the highest average accuracy across all training-test splits in the bottom row.

of the features, models and approaches used, accuracy generally reaches around 70%, and it is difficult to increase beyond that level. The default number of boosting iterations in WEKA, each of which creates an additional ADTree branch, is 10. The ADTree that achieved the equal-highest accuracy across the test sets (72.7%; training set 2006-2014, test set 2015), shown in Figure 2, was overly complex since the AvgOddsDiff feature was present in multiple branches.

However, for the 2006-2014 training 2015 test split, even when the number of boosting iterations was set to one, that is, the ADTree has only one branch (Figure 3), the same accuracy as the 10-branch ADTree was achieved. Since the sum of each predictor node in the ADTree determines the class in which a fresh instance is to be classified, the decision rules that can be derived from this ADTree are:

```
if AvgOddsDiff>=-0.431, 0.014+(-0.401)=-0.387<0
    => P2 result = "W";
else if AvgOddsDiff<-0.431, 0.014+0.509=0.523>0
    => P2 result = "L"
```

Likewise, Figure 4 shows an ADTree with one boosting iteration, trained on 2006-2008 matches and tested on 2009 matches, which achieved 71.2% accuracy, which was the same as the 10-boosting-iteration ADTree and higher than the 70.8% accuracy that was derived from the betting odds. The decision rules that can be derived from this ADTree are:

```
if AvgOddsDiff>=0.328, 0.023+(-0.515)=-0.492<0
    => P2 result = "W";
else if AvgOddsDiff<0.328, 0.023+0.389=0.412>0
    => P2 result = "L"
```

Figure 2. ADTree with 10 boosting iterations, trained on 2006-2014 matches and tested on 2015 matches, which achieved the equal-highest accuracy of 72.7%.

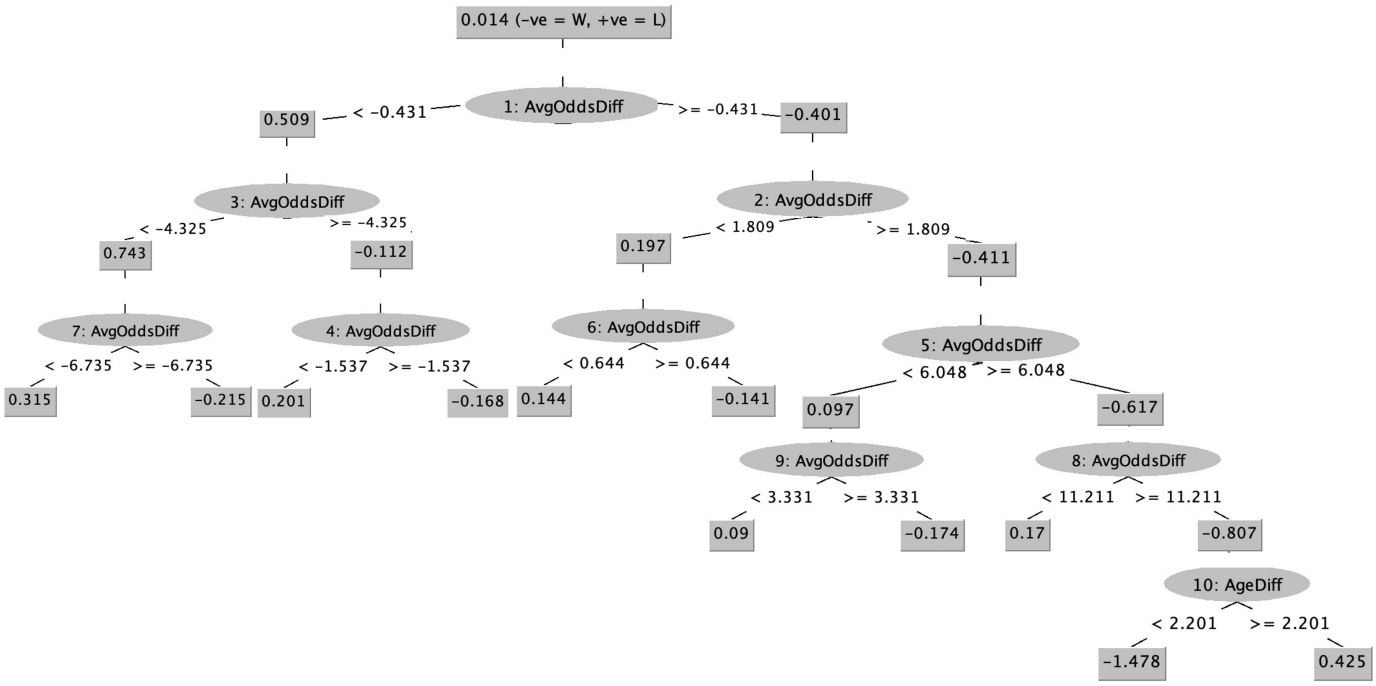
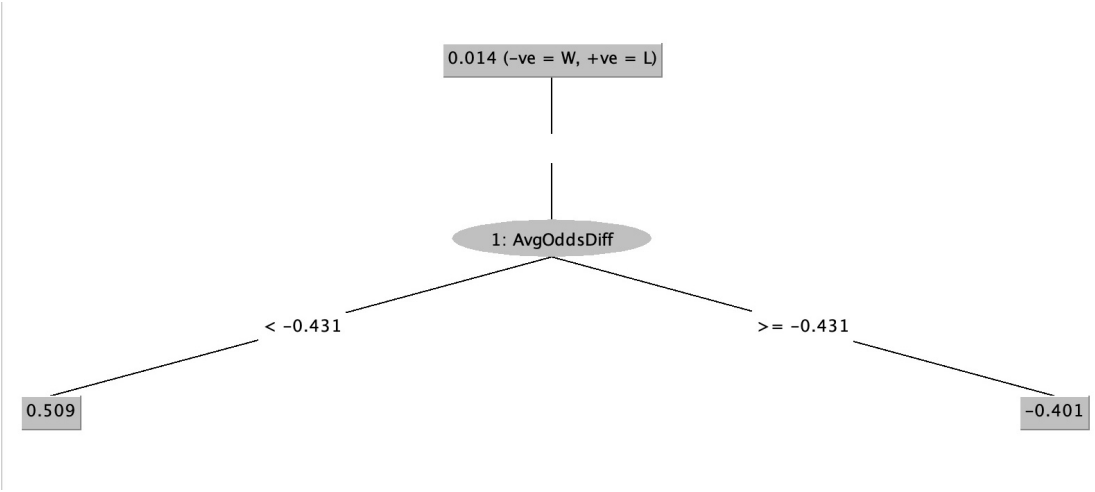


Figure 3. ADTree with one boosting iteration, trained on 2006-2014 matches and tested on 2015 matches, which also achieved 72.7% accuracy.



In fact, the one-branch ADTrees provided the same accuracy performance as the ten-branch ADTrees in all but one test set, the 2020 test set, where the one-branch ADTree achieved a slightly lower accuracy of 68.3%. While predictions derived from betting odds use an AvgOddsDiff threshold of zero on the test set to determine the target variable value, the trained ADTrees allow for this threshold to vary across test sets. The AvgOddsDiff thresholds for each training-test split, which split the one-branch ADTrees, are shown in Table 4.

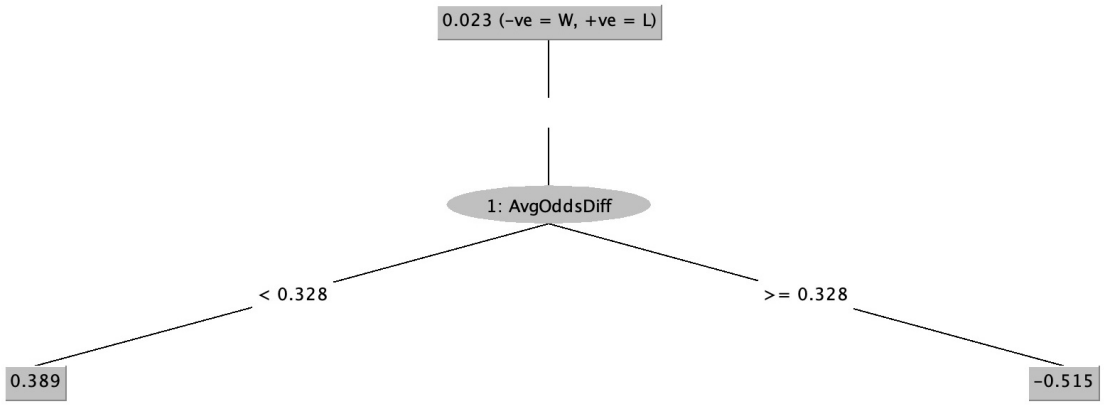
As can be seen, while deriving predictions purely from the betting odds implies a threshold value of zero for AvgOddsDiff, the ADTree allows for this threshold to vary, for example, to -0.431 in the case of the ADTree in Figure 3 and 0.328 in the case of the ADTree in Figure 4.

Table 5 shows the differences in accuracy between each of the ML models and Elo/WElo. The results show that, unlike the ANN, SVM and Random Forest models, the ADTree and Logistic Regression models outperformed both Elo and WElo

Table 4. AvgOddsDiff thresholds from the one-branch ADTrees for each training-test split. All 1-branch ADTrees, apart from this training-test split (where the 1-branch ADTree achieved accuracy of 68.3% versus the 10-branch tree's 69.8%), obtained the same accuracy as the default 10-branch ADTrees.

Train	Test	#Train	#Test	Train%	Test%	ADTree Accuracy (1-branch)	AvgOddsDiff Threshold
2006	2007	2,306	2,368	49%	51%	71.0	0.312
2006-2007	2008	4,674	2,242	68%	32%	70.9	0.328
2006-2008	2009	6,916	2,278	75%	25%	71.2	-0.431
2006-2009	2010	9,194	2,350	80%	20%	63.8	0.328
2006-2010	2011	11,544	2,345	83%	17%	71.4	0.328
2006-2011	2012	13,889	2,324	86%	14%	72.6	0.328
2006-2012	2013	16,213	2,316	88%	12%	68.9	0.328
2006-2013	2014	18,529	2,250	89%	11%	70.9	-0.431
2006-2014	2015	20,779	2,285	90%	10%	72.7	-0.431
2006-2015	2016	23,064	2,297	91%	9%	70.3	-0.431
2006-2016	2017	25,361	2,307	92%	8%	68.0	0.050
2006-2017	2018	27,668	2,008	93%	7%	67.9	-0.117
2006-2018	2019	29,676	2,303	93%	7%	67.3	-0.431
2006-2019	2020	31,979	1,033	97%	3%	68.3	-0.431*

Figure 4. ADTree with one boosting iteration, trained on 2006-2008 matches and tested on 2009 matches, which achieved 71.2% accuracy (the same as the 10-boosting-iteration ADTree, and higher than the 70.8% accuracy derived from the betting odds in the 2009 test set).



on all training-test splits. The respective average differences in accuracy between the ML models and Elo/WElo across all training-test splits were 2.3% and 2.0% for ADTree and Logistic Regression. It is also notable that Elo and WElo ratings tended to outperform the SVM when the amount of training data used was small.

Table 5. Accuracy differences (%) between the ML models and Elo/WElo

Train	Test	#Train	#Test	Train%	Test%	LR-Elo	LR-WElo	ANN-Elo	ANN-WElo	SVM-Elo	SVM-WElo	RF-Elo	RF-WElo	ADT-Elo	ADT-WElo
2006	2007	2,306	2,368	49%	51%	3.6	4.5	-2.2	-1.3	-3.6	-2.8	1.5	2.4	3.7	4.6
2006-2007	2008	4,674	2,242	68%	32%	2.8	3.5	-2.3	-1.6	-3.5	-2.8	-1.3	-0.6	3.4	4.1
2006-2008	2009	6,916	2,278	75%	25%	0.7	0.7	-3.8	-3.8	-3.1	-3.1	-1.9	-1.8	1.5	1.5
2006-2009	2010	9,194	2,350	80%	20%	-5.8	-6.4	-4.6	-5.1	-3.4	-4.0	-0.6	-1.1	-4.2	-4.8
2006-2010	2011	11,544	2,345	83%	17%	2.3	2.3	0.0	0.0	-0.7	-0.7	-2.7	-2.7	2.1	2.1
2006-2011	2012	13,889	2,324	86%	14%	0.9	1.4	-1.5	-1.0	-3.1	-2.6	-2.3	-1.8	1.5	1.9
2006-2012	2013	16,213	2,316	88%	12%	1.9	1.0	0.5	-0.4	0.3	-0.6	-0.8	-1.7	1.6	0.7
2006-2013	2014	18,529	2,250	89%	11%	2.2	2.4	-0.3	-0.1	0.8	1.0	-2.0	-1.7	2.0	2.2
2006-2014	2015	20,779	2,285	90%	10%	2.4	2.5	-3.3	-3.2	2.0	2.1	1.1	1.1	3.3	3.4
2006-2015	2016	23,064	2,297	91%	9%	2.8	2.4	2.0	1.6	1.7	1.3	-0.9	-1.3	1.7	1.3
2006-2016	2017	25,361	2,307	92%	8%	0.8	1.3	-1.2	-0.7	-0.3	0.1	-0.6	-0.1	1.0	1.5
2006-2017	2018	27,668	2,008	93%	7%	4.3	3.7	3.6	3.0	2.8	2.2	0.2	-0.4	4.6	4.0
2006-2018	2019	29,676	2,303	93%	7%	3.1	2.9	2.5	2.2	1.8	1.6	0.9	0.6	3.6	3.3
2006-2019	2020	31,979	1,033	97%	3%	5.8	5.3	3.9	3.4	6.0	5.5	1.3	0.8	6.2	5.7
						2.0	2.0	-0.5	-0.5	-0.2	-0.2	-0.6	-0.6	2.3	2.3

Table 6 shows the differences in accuracy between the ML models and odds-derived predictions on the test sets. The results show that ADTree and Logistic Regression performed comparably to betting odds-derived predictions; however, ANN, SVM, Random Forest, and Elo/WElo performed worse than betting odds.

Table 6. Accuracy differences (%) between the ML models and betting odds-derived predictions on the test sets.

Test	#Train	#Test	Train%	Test%	LR-Odds	ANN-Odds	SVM-Odds	RF-Odds	ADT-Odds	Elo-Odds	WElo-Odds
2007	2,306	2,368	49%	51%	-0.6	-6.4	-7.9	-2.7	-0.5	-4.2	-5.1
2008	4,674	2,242	68%	32%	-0.5	-5.6	-6.9	-4.6	0.1	-3.3	-4.1
2009	6,916	2,278	75%	25%	-0.5	-5.0	-4.3	-3.0	0.4	-1.1	-1.2
2010	9,194	2,350	80%	20%	-8.6	-7.3	-6.2	-3.4	-7.0	-2.8	-2.2
2011	11,544	2,345	83%	17%	-0.8	-3.1	-3.7	-5.7	-0.9	-3.0	-3.0
2012	13,889	2,324	86%	14%	-0.6	-3.0	-4.6	-3.8	0.0	-1.5	-2.0
2013	16,213	2,316	88%	12%	-0.2	-1.6	-1.8	-2.8	-0.5	-2.1	-1.2
2014	18,529	2,250	89%	11%	-0.3	-2.8	-1.7	-4.4	-0.4	-2.4	-2.7
2015	20,779	2,285	90%	10%	-0.7	-6.4	-1.1	-2.1	0.2	-3.1	-3.2
2016	23,064	2,297	91%	9%	0.3	-0.5	-0.8	-3.4	-0.8	-2.5	-2.1
2017	25,361	2,307	92%	8%	-0.1	-2.1	-1.3	-1.5	0.1	-0.9	-1.4
2018	27,668	2,008	93%	7%	0.0	-0.7	-1.5	-4.1	0.3	-4.3	-3.7
2019	29,676	2,303	93%	7%	0.0	-0.7	-1.3	-2.3	0.4	-3.2	-2.9
2020	31,979	1,033	97%	3%	-0.3	-2.2	-0.1	-4.8	0.1	-6.1	-5.6
					-0.9	-3.4	-3.1	-3.5	-0.6	-2.9	-2.9

Conclusions

In this study, the performance in terms of accuracy of five machine learning (ML) models was compared with Elo and WElo ratings in predicting the results of professional men's ATP tennis matches. The differences in performance between the ML models and Elo/WElo and betting odds-derived predictions were also examined.

The results showed that across the 14 training and test sets considered, Logistic Regression and ADTree achieved an average accuracy of 69.5% and 69.8%, respectively, performing comparably to betting odds-derived predictions; however, the ANN, SVM, and Random Forest models – with respective average accuracies of 67%, 67.3%, and 66.9%, and Elo and WElo, which both had an average accuracy of 67.5% – performed worse. Furthermore, the Logistic Regression and ADTree models outperformed Elo and WElo on all training-test splits except for one that was afflicted by a high proportion of null values in the average betting odds difference feature. The ADTree model, by allowing the decision threshold for the average betting odds difference to vary across training-test splits, outperformed odds-derived predictions in half of the training-test splits considered. The dominant model feature was the average (across 11 betting companies) betting odds difference. Since ADTrees retain the interpretability of a decision tree structure, as opposed to commonly used boosted tree methods such as XGBoost and CatBoost, ADTrees may have potential in some areas of sports analytics.

Our results were consistent with recent studies[15,50–52] that have found that accuracy in professional men's tennis match result prediction tends to reach approximately 70%, which is difficult to improve upon and is similar to the accuracy achieved by betting odds alone. This work has added to the growing body of literature suggesting that boosting methods perform well in predicting match results in sports,[51,53,54] and features that contain aggregated information including historical match result results, for example, ratings or betting odds (which can also be considered a rating[55]), can be important.

In future research, historically averaged in-play features could be included as features, the dominant betting odds features could be removed for comparison, and additional (tuned) boosting methods and alternative ratings could be added to the comparison.

Supplemental material

R Code. The R script that was used to apply Elo and WElo is available on GitHub (https://github.com/rorybunker/tennis-elo-vs-ml).

Data Availability. The machine learning models were applied in WEKA and the training/test set (ARFF) files are also available in the above GitHub repository. As mentioned in the Data Collection subsection, the ATP RData file is openly available in the appendix of the online version of the paper of Angelini, Candila, & De Angelis[13] at https://doi.org/10.1016/j.ejor.2021.04.011 under "Supplementary Data S1". This is open data under the CC BY license http://creativecommons.org/licenses/by/4.0/.

In-play features statistical analysis. As mentioned, the in-play variables could not be used for match result prediction since the values of these variables are only known post-match. However, out of practical interest, statistical techniques were used to analyse the differences in these variables between winning and losing outcomes (from Player 2's perspective).

As shown by the Anderson-Darling normality test p-values, which were all very close to zero, the null hypothesis of normality for each of the six features was rejected. Because the two-sample t-test was unable to be employed, the non-parametric Wilcoxon-Mann-Whitney test was used. All p-values from the Wilcoxon-Mann-Whitney test were very close to zero, indicating statistically significant differences between the win and loss class labels for all six variables. Cohen's D effect sizes showed that the difference in the percentage of points won on a first serve ($D = 1.9$), followed by the difference in the percentage of breakpoints saved ($D = 0.91$), followed by the difference in the percentage of first serves that were aces ($D = 0.8$) were the most relevant discriminative features.

Feature ranks for each feature selection technique. Here, in Tables 8 to 12, the feature ranks for each of the feature selection techniques considered are presented, which were used to calculate the average feature ranks in Table 2 and then to select the final feature set.

Table 7. Descriptive statistics between the "W" and "L" target variable labels for the in-play features in the transformed dataset, along with the p-values from Anderson-Darling test for normality (shown in column pA) and Wilcoxon-Mann-Whitney's test p-values (shown in column pW), as well as Cohen's D effect sizes (shown in column D). The Wilcoxon-Mann-Whitney test p-values were all <2.2E-16, and are displayed in the table below as ˜0.

target	feature	N	mean	SD	med	min	max	range	skew	kurt	pA	pW	D
L	AcePercDiff	17072	0.03	0.08	0.03	-0.39	0.53	0.91	0.44	1.63	3.7E-24	˜0	0.8
	DFPercDiff	17072	-0.01	0.04	-0.01	-0.23	0.19	0.41	-0.26	1.00	3.7E-24	˜0	0.5
	FirstServePercDiff	17072	0.02	0.11	0.02	-0.43	0.49	0.92	0.07	0.19	7.8E-06	˜0	0.4
	FirstServePtWinPerc	17072	0.11	0.11	0.10	-0.27	0.71	0.98	0.46	0.53	2.9E-04	˜0	1.9
	PtsWonOnScnd ServePerService GameDiff	16832	0.13	0.48	0.12	-3.50	2.5	6.00	0.11	0.39	9.4E-14	˜0	0.5
	BreakPtSavePerc	15409	0.15	0.33	0.16	-1.00	1.00	2.00	-0.18	0.42	1.2E-19	˜0	0.9
W	AcePercDiff	16775	-0.03	0.08	-0.02	-0.47	0.32	0.78	-0.41	1.41			
	DFPercDiff	16775	0.01	0.04	0.01	-0.16	0.27	0.44	0.33	1.36			
	FirstServePercDiff	16775	-0.02	0.11	-0.02	-0.45	0.38	0.83	-0.02	0.18			
	FirstServePtWinPerc	16775	-0.10	0.11	-0.09	-0.67	0.26	0.93	-0.44	0.42			
	PtsWonOnScnd ServePerService GameDiff	16536	-0.12	0.48	-0.11	-2.67	2.00	4.67	-0.08	0.23			
	BreakPtSavePerc	15189	-0.15	0.33	-0.17	-1.00	1.00	2.00	0.19	0.34			

Table 8. Feature ranks in terms of correlation

Rank	Correlation	Feature
1	0.406	AvgOddsDiff
2	0.395	B365OddsDiff
3	0.329	PtsDiffNorm
4	0.329	PtsDiff
5	0.285	RankDiffNorm
6	0.285	RankDiff
7	0.237	PtsDiffPerc
8	0.191	RankDiffPerc
9	0.168	P1_Seed
10	0.096	P1_Entry
11	0.059	HeightDiff
12	0.025	SameHand
13	0.022	AgeDiff
14	0.003	SameCountry

Table 9. Feature ranks in terms of Chi-squared statistic

Rank	ChiSquared	Feature
1	7831.845	B365OddsDiff
2	7475.673	AvgOddsDiff
3	5191.824	RankDiffPerc
4	5058.638	PtsDiffPerc
5	5054.41	PtsDiffNorm
6	5054.41	PtsDiff
7	4066.51	RankDiffNorm
8	4066.51	RankDiff
9	2284.348	P1_Seed
10	368.704	P1_Entry
11	97.871	HeightDiff
12	21.176	SameHand
13	0.376	SameCountry
14	0	AgeDiff

Table 10. Feature ranks in terms of information gain ratio

Rank	GainRatio	Feature
1	0.0516943	AvgOddsDiff
2	0.0505759	B365OddsDiff
3	0.0389762	RankDiffPerc
4	0.0385497	PtsDiffPerc
5	0.0358154	PtsDiffNorm
6	0.0358154	PtsDiff
7	0.0302545	RankDiffNorm
8	0.0302545	RankDiff
9	0.0193729	P1_Seed
10	0.0091313	P1_Entry
11	0.0013614	HeightDiff
12	0.0005692	SameHand
13	0.0000213	SameCountry
14	0	AgeDiff

Table 11. Feature ranks in terms of symmetrical uncertainty

Rank	Feature	SymmU
1	AvgOddsDiff	0.0798728
2	B365OddsDiff	0.0794467
3	RankDiffPerc	0.0587877
4	PtsDiffPerc	0.0577436
5	PtsDiffNorm	0.0546706
6	PtsDiff	0.0546706
7	RankDiffNorm	0.0451663
8	RankDiff	0.0451663
9	P1_Seed	0.0279571
10	P1_Entry	0.0084697
11	HeightDiff	0.001646
12	SameHand	0.0005024
13	SameCountry	0.0000116
14	AgeDiff	0

Table 12. Feature ranks in terms of ReliefF attribute evaluation

Feature	ReliefF	Rank
AvgOddsDiff	0.0048062	1
PtsDiffNorm	0.0040838	2
PtsDiff	0.0040838	3
B365OddsDiff	0.0036873	4
AgeDiff	0.0034504	5
HeightDiff	0.0031599	6
RankDiff	0.0025382	7
RankDiffNorm	0.0025382	8
RankDiffPerc	0.0011072	9
PtsDiffPerc	0.0009	10
P1_Seed	0.0008418	11
SameHand	0.0001295	12
SameCountry	0.0001295	13
P1_Entry	-0.0000294	14

Funding

This work was partly supported by JSPS KAKENHI (grant number 20H04075) and JST Presto (grant number JPMJPR20CA).

Declaration of conflicting interests

The authors declare that there are no conflicts of interest.

References

1. Lake RJ, editor. Routledge handbook of tennis: History, culture and politics. *Routledge, Abingdon, U.K.*; 2019 Feb 5.
2. International Tennis Federation. ITF Global Tennis Report 2019: A Report on Tennis Participation and Performance Worldwide [Internet]. 2019 [cited 2023 Jul 12]. Available from: http://itf.uberflip.com/i/1169625-itf-global-tennis-report-2019-overview/0
3. Kovalchik, S. Searching for the GOAT of tennis win prediction. *Journal Of Quantitative Analysis In Sports*. **12**, 127-138 (2016)
4. Kovalchik, S. Extension of the Elo rating system to margin of victory. *International Journal Of Forecasting*. **36**, 1329-1341 (2020)
5. Elo, A. The rating of chess players, past and present. (Arco Pub., New York, 1978)
6. Hvattum, L. & Arntzen, H. Using Elo ratings for match result prediction in association football. *International Journal Of Forecasting*. **26**, 460-470 (2010)
7. Ryall, R. & Bedford, A. An optimized ratings-based model for forecasting Australian Rules football. *International Journal Of Forecasting*. **26**, 511-517 (2010)
8. Leitner, C., Zeileis, A. & Hornik, K. Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the EURO 2008. *International Journal Of Forecasting*. **26**, 471-481 (2010)
9. Constantinou, A. & Fenton, N. Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal Of Quantitative Analysis In Sports*. **9**, 37-50 (2013)
10. Van Haaren, J. & Davis, J. Predicting the final league tables of domestic football leagues. *Proceedings Of The 5th International Conference On Mathematics In Sport*. pp. 202-207 (2015). Loughborough, U.K. 29 June - 1 July 2015.
11. Mangan, S. & Collins, K. A rating system for Gaelic football teams: Factors that influence success. *Journal Homepage: Http://iacss. Org/index. Php? Id*. **15** (2016) pp. 78 - 90.
12. Robberechts, P. & Davis, J. Forecasting the FIFA World Cup–Combining result-and goal-based team ability parameters. *Machine Learning and Data Mining for Sports Analytics: 5th International Workshop, MLSA 2018, Co-located with ECML/PKDD 2018, Dublin, Ireland, September 10*. pp. 16-30 (2018)
13. Angelini, G., Candila, V. & De Angelis, L. Weighted Elo rating for tennis match predictions. *European Journal Of Operational Research*. https://doi.org/10.1016/j.ejor.2021.04.011. (2021)
14. Williams, L., Liu, C., Dixon, L. & Gerrard, H. How well do Elo-based ratings predict professional tennis matches?. *Journal Of Quantitative Analysis In Sports*. **17**, 91-105 (2021)
15. Wilkens, S. Sports prediction and betting models in the machine learning age: The case of tennis. *Journal Of Sports Analytics*. **7**, 99-117 (2021)
16. Freund, Y. & Mason, L. The alternating decision tree learning algorithm. *Icml*. **99** pp. 124-133 (1999)
17. Pfahringer, B., Holmes, G. & Kirkby, R. Optimizing the induction of alternating decision trees. *Pacific-Asia Conference On Knowledge Discovery And Data Mining, Hong Kong, China*. pp. 477-487 (2001)
18. Comité, F., Gilleron, R. & Tommasi, M. Learning multi-label alternating decision trees from texts and data. *International Workshop On Machine Learning And Data Mining In Pattern Recognition*. pp. 35-49 (2003)
19. Jabbar, M., Deekshatulu, B. & Chndra, P. Alternating decision trees for early diagnosis of heart disease. *International Conference On Circuits, Communication, Control And Computing*. pp. 322-328 (2014)
20. Liu, K., Lin, J., Zhou, X. & Wong, S. Boosting alternating decision trees modeling of disease trait information. *BMC Genetics*. **6**, 1-6 (2005)
21. Pham, B., Tien Bui, D. & Prakash, I. Landslide susceptibility assessment using bagging ensemble based alternating decision trees, logistic regression and J48 decision trees methods: a comparative study. *Geotechnical And Geological Engineering*. **35**, 2597-2611 (2017)
22. Hong, H., Liu, J. & Zhu, A. Modeling landslide susceptibility using LogitBoost alternating decision trees and forest by penalizing attributes with the bagging ensemble. *Science Of The Total Environment*. **718** pp. 137231 (2020)

23. Costache, R., Arabameri, A., Blaschke, T., Pham, Q., Pham, B., Pandey, M., Arora, A., Linh, N. & Costache, I. Flash-flood potential mapping using deep learning, alternating decision trees and data provided by remote sensing sensors. *Sensors*. **21**, 280 (2021)
24. Bunker, R. & Thabtah, F. A machine learning framework for sport result prediction. *Applied Computing And Informatics*. **15**, 27-33 (2019)
25. Clarke, S. & Dyte, D. Using official ratings to simulate major tennis tournaments. *International Transactions In Operational Research*. **7**, 585-594 (2000)
26. Klaassen, F. & Magnus, J. Forecasting the winner of a tennis match. *European Journal Of Operational Research*. **148**, 257-267 (2003)
27. Lisi, F. Tennis betting: can statistics beat bookmakers?. *Electronic Journal Of Applied Statistical Analysis*. **10**, 790-808 (2017)
28. Makino, M., Odaka, T., Kuroiwa, J., Suwa, I. & Shirai, H. Feature Selection to Win the Point of ATP Tennis Players Using Rally Information.. *Int. J. Comput. Sci. Sport*. **19**, 37-50 (2020)
29. Ghosh, S., Sadhu, S., Biswas, S., Sarkar, D. & Sarkar, P. A comparison between different classifiers for tennis match result prediction. *Malaysian Journal Of Computer Science*. **32**, 97-111 (2019)
30. Gu, W. & Saaty, T. Predicting the outcome of a tennis tournament: Based on both data and judgments. *Journal Of Systems Science And Systems Engineering*. **28**, 317-343 (2019)
31. Saaty, T. Decision making with dependence and feedback: The analytic network process. (RWS publications Pittsburgh,1996)
32. Candila, V. & Palazzo, L. Neural networks and betting strategies for tennis. *Risks*. **8**, 68 (2020)
33. Gao, Z. & Kowalczyk, A. Random forest model identifies serve strength as a key predictor of tennis match outcome. *Journal Of Sports Analytics*. **7**, 255-262 (2021)
34. Breiman, L. Random forests. *Machine Learning*. **45**, 5-32 (2001)
35. Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning*. **20**, 273-297 (1995)
36. Del Corral, J. & Prieto-Rodriguez, J. Are differences in ranks good predictors for Grand Slam tennis matches?. *International Journal Of Forecasting*. **26**, 551-563 (2010)
37. Bradley, R. & Terry, M. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*. **39**, 324-345 (1952)
38. Coulom, R. Computing "Elo ratings" of move patterns in the game of go. *ICGA Journal*. **30**, 198-208 (2007)
39. McHale, I. & Morton, A. A Bradley-Terry type model for forecasting tennis match results. *International Journal Of Forecasting*. **27**, 619-630 (2011)
40. Baker, R. & McHale, I. A dynamic paired comparisons model: Who is the greatest tennis player?. *European Journal Of Operational Research*. **236**, 677-684 (2014)
41. Fayomi, A., Majeed, R., Algarni, A., Akhtar, S., Jamal, F. & Nasir, J. Forecasting Tennis Match Results Using the Bradley-Terry Model. *International Journal Of Photoenergy*. **2022** (2022)
42. Temesi, J., Szádoczki, Z. & Bozóki, S. Incomplete pairwise comparison matrices: Ranking top women tennis players. *Journal Of The Operational Research Society*. pp. 1-13 (2023)
43. Han, R., Ye, R., Tan, C. & Chen, K. Asymptotic theory of sparse Bradley–Terry model. *Annals Of Applied Probability*. pp. 2491 (2020)
44. Morris, B., Bialik, C. & Boice, J. How We're Forecasting The US Open. [Internet]. New York (US): FiveThirtyEight (ABC News Ventures); 2016 Aug 28 [cited 2023 Sep 14]. Available from: https://fivethirtyeight.com/features/how-were-forecasting-the-2016-us-open/
45. Candila V (2022). *WElo: Weighted and Standard Elo Rates*. R package version 0.1.3.
46. Wirth, R. & Hipp, J. CRISP-DM: Towards a standard process model for data mining. *Proceedings Of The 4th International Conference On The Practical Applications Of Knowledge Discovery And Data Mining*. **1** pp. 29-39 (2000)
47. Tax, N. & Joustra, Y. Predicting the Dutch football competition using public data: A machine learning approach. *Transactions On Knowledge And Data Engineering*. **10**, 1-13 (2015)
48. Hubáček, O., Šourek, G. & Železnỳ, F. Exploiting sports-betting market using machine learning. *International Journal Of Forecasting*. **35**, 783-796 (2019)
49. Kira, K. & Rendell, L. A practical approach to feature selection. *Machine Learning Proceedings 1992*. pp. 249-256 (1992)
50. De Araujo Fernandes M. Using soft computing techniques for prediction of winners in tennis matches. Machine Learning Research. 2017;2(3):86-98.

51. Chavda J, Patel N, Vishwakarma P. Predicting tennis match winner and comparing bookmakers odds using machine learning techniques. National College of Ireland, Dublin. 2019.
52. Cornman A, Spellman G, Wright D. Machine learning for professional tennis match prediction and betting. Working Paper, Stanford University. 2017.
53. Razali MN, Mustapha A, Mostafa SA, Gunasekaran SS. Football Matches Outcomes Prediction Based on Gradient Boosting Algorithms and Football Rating System. Human Factors in Software and Systems Engineering. 2022 Jul 24;61:57.
54. Dubitzky W, Lopes P, Davis J, Berrar D. The open international soccer database for machine learning. Machine learning. 2019 Jan 15;108:9-28.
55. Wunderlich F, Memmert D. The betting odds rating system: Using soccer forecasts to forecast soccer. PloS one. 2018 Jun 5;13(6):e0198668.

CHAPTER 9: PUBLICATION 6 - “MACHINE LEARNING FOR SOCCER MATCH RESULT PREDICTION” (BOOK CHAPTER)

This chapter provided a comprehensive survey and synthesis in machine learning for soccer outcome prediction (i.e., performance was considered at the match level). Soccer was investigated not only because of its global popularity but also because it was identified by Bunker and Susnjak (2022) as a sport with inherent characteristics that make it challenging to predict. Available datasets, models, features, and model performance evaluation approaches for machine learning for match outcome prediction in soccer were outlined. Model features were discussed in detail, including those derived from historical match outcomes, team/player ratings, competition standings, spatiotemporal tracking data, and event log data. State-of-the-art approaches in the field, including gradient-boosted tree models such as CatBoost and XGBoost applied to rating features, were also discussed in the chapter, as were random forests, deep learning, and traditional statistical techniques. In addition, model performance evaluation metrics such as classification accuracy, the Brier score, the ranked probability score, root mean squared error, and the ignorance score were outlined. The appropriateness of specific metrics in different problem settings, for example, based on differing target variable definitions (numeric or discrete target) and dataset imbalance, was also analysed.

As avenues for future work, we suggested that comparing gradient-boosted tree models with the performance of newer deep learning models and random forests on a range of datasets with different types of features is necessary to establish the state-of-the-art in the domain. It was also suggested that the interpretability of match result prediction models be enhanced to have greater utility for coaches and performance analysts. Furthermore, it was noted that new rating systems that utilise both player- and team-level information and that incorporate additional information from spatiotemporal tracking data and event log data could be utilised as model features. By providing a thorough overview of current techniques and discussing promising avenues for future research, this chapter acts as a valuable resource for researchers in the domain.

Bunker, R., Yeung, C., Fujii, K. (2025). Machine Learning for Soccer Match Result Prediction. In: Blondin, M.J., Fister Jr., I., Pardalos, P.M. (eds) *Artificial Intelligence, Optimization, and Data Sciences in Sports. Springer Optimization and Its Applications*, vol 218. Springer, Cham. https://doi.org/10.1007/978-3-031-76047-1_2

Machine Learning for Soccer Match Result Prediction

Rory Bunker, Calvin Yeung, and Keisuke Fujii

Abstract Machine learning has become a common approach to predicting the outcomes of soccer matches, and the body of literature in this domain has grown substantially in the past decade and a half. This chapter discusses available datasets, the types of models and features, and ways of evaluating model performance in this application domain. The aim of this chapter is to give a broad overview of the current state and potential future developments in machine learning for soccer match results prediction, as a resource for those interested in conducting future studies in the area. Our main findings are that while gradient-boosted tree models such as CatBoost, applied to soccer-specific ratings such as pi-ratings, are currently the best-performing models on datasets containing only goals as the match features, there needs to be a more thorough comparison of the performance of deep learning models and Random Forest on a range of datasets with different types of features. Furthermore, new rating systems using both player- and team-level information and incorporating additional information from, e.g., spatiotemporal tracking and event data, could be investigated further. Finally, the interpretability of match result prediction models needs to be enhanced for them to be more useful for team management.

Rory Bunker
Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa ward, Nagoya, 464-8601, Japan, e-mail: rory.bunker@g.sp.m.is.nagoya-u.ac.jp

Calvin Yeung
Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa ward, Nagoya, 464-8601, Japan, e-mail: yeung.chikwong@g.sp.m.is.nagoya-u.ac.jp

Keisuke Fujii
Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa ward, Nagoya, 464-8601, Japan, e-mail: fujii@i.nagoya-u.ac.jp

1 Introduction

Predicting the results of professional soccer[1] matches is a challenging problem due to draws being a common outcome in the sport, as well as its low-scoring nature and often highly competitive leagues. Nonetheless, given the global popularity of the sport, both in terms of spectatorship and player numbers, it is a topic that is of interest to many groups including fans, bookmakers and bettors, as well as coaches, players, and performance analysts. While bettors and bookmakers require models that are highly accurate, coaches/management and sports performance analysts also require models that are interpretable, so that the most relevant match features (performance indicators [63]) for winning can be identified and improved upon in future matches. Once a predicted result for a specific match is obtained, an additional problem is to decide whether to actually bet on the match. While this is an important question [38], it is not the focus of the current chapter.

A large number of papers have been published that are related to machine learning (ML) for soccer match result prediction, particularly over the past decade [18]. Traditionally, however, statistical models were used to forecast soccer match results. Stefani [105] used least-squares regression to calculate team ratings, which could be updated on a weekly basis, to predict match results based on the difference in ratings between teams. Some early papers fit distributions to the number of goals scored by each team in a match. Maher [81] used an independent Poisson distribution to obtain the attacking and defensive ratings of teams. Dixon & Coles [37] modified this model to handle incomplete data and data from different divisions, and to allow for temporal variations in the performance of teams. Goals distributions have also been fit using, e.g., the dependent Poisson, negative binomial, and extreme value distributions [51, 9, 10, 91]. Indeed, statistical models such as the Bivariate Poisson still provide strong performance (e.g., [78] who used the engsoccerdata package [34]).

Constantinou [28] categorized soccer result prediction models into three groups, suggesting that the first and third types of studies tend to be published in statistical journals, while the second group tend to be published in journals focused on computer science and artificial intelligence.

1. **Statistical models**: Including modeling goals scored, e.g., using the Bivariate, independent and dependent Poisson, negative binomial, and generalized extreme value distributions [81, 37, 51, 9, 10, 91], as well as ordered logistic regression models [4, 100].
2. **Machine learning and probabilistic graphical models**: Including fuzzy or genetic algorithms [112, 87, 54, 95] and Bayesian methods (e.g., the studies of Joseph, Fenton & Neil [68] and Constantinou, Fenton, & Neil [31]).
3. **Rating systems**: Including Elo ratings [41, 65], pi-ratings [30], and more recently, Berrar ratings [8] and Generalized Attacking Performance (GAP) ratings [118]. PageRank ratings [15, 59] also fit under this category.[2]

[1] Also known as "Association football" or simply "football".

[2] How these different ratings are calculated is described in subsection 4.1.

Nonetheless, this three-group model taxonomy is not entirely satisfactory. In recent years, researchers have incorporated models, techniques, or features from more than one of the above three groups to create hybrid approaches. For instance, the top-performing participants in the 2017 Soccer Prediction Challenge [7] applied machine learning models [8, 59, 28] to soccer-specific ratings. Statistical and machine learning models have also been combined to form hybrid approaches [40, 74, 53], with [53], for example, proposing a hybrid approach combining a Random Forest model with Poisson-based rankings. In addition, some models, e.g., Logistic Regression, fit under both the umbrella of statistics as well as that of machine learning. On the other hand, Bradley-Terry models [11], which also proved successful during the 2017 Soccer Prediction Challenge [116] and for ranking soccer teams based on current strength [78], could be considered to fit under both the statistical models and rating systems groups. This is because, as will be explored later in this chapter in subsection 4.1.1, Bradley-Terry forms the theoretical foundation for Elo ratings [33].

By covering available datasets — including an in-depth discussion of the Open International Database and the 2017 Soccer Prediction Challenge — as well as current and potential future models and features, as well as evaluation methods, this chapter aims to provide a broad overview of machine learning for soccer match result prediction and will hopefully act as a resource for those interested in carrying out future research in the domain.

The remainder of this chapter is organized as follows. In the next section, available datasets are outlined, as are the Open International Soccer Database and 2017 Soccer Prediction Challenge's in-competition and post-competition approaches and results. Subsequently, commonly used candidate models including conventional machine learning models, ensemble methods, and deep learning models are considered, as are model objectives and interpretability. Then, current and potential future model features are considered, e.g., match features, player and team statistics, and ratings (both general and soccer-specific). Feature selection methods are also briefly discussed. Model evaluation considerations — including baselines, target variable definition, as well as scoring rules and their properties, along with temporal splitting methods — are then described. Finally, conclusions and potential avenues for future research are discussed.

2 Data

In subsection 2.1, a non-exhaustive list of the datasets that are available for engineering features and building models is described. The dataset used in the 2017 Soccer Prediction Challenge [7], the Open International Soccer Database [39], will then be discussed, which — despite not containing match features derived from in-play events during games (apart from goals scored) — now acts as somewhat of a benchmark dataset. The top-performing in-competition and post-competition studies that have used the Open International Soccer Database will be discussed in subsection 2.2.

2.1 Available Datasets

An increasing amount of publicly available data from soccer matches is becoming publicly available online. Organizations such as StatsBomb (Bath, United Kingdom) make some event log data publicly available, although the data is generally held by professional teams who pay vendors such as StatsBomb or Stats Perform (Chicago, IL, USA)/Opta (London, United Kingdom) for access. A non-exhaustive list of datasets that can be used for engineering features and building soccer match result prediction ML models is presented in Table 1.

One challenge with many of these data sources is that they often cover different leagues and/or different seasons, and may be missing some types of features. The Open International Soccer Database, for instance, does not contain match features apart from goals scored, nor does it contain betting odds. It may be necessary for researchers to source data from different datasets and merge them. The European Soccer Database on Kaggle has already performed such merging for a variety of feature types for a subset of their entire dataset (10,000 of the 25,000 matches). Spatiotemporal data is often only available to professional teams themselves, although sometimes small amounts of it are made publicly available.

2.2 2017 Soccer Prediction Challenge

For the 2017 Open International Soccer Prediction Challenge, the Open International Soccer Database [39], an open-source database containing over 216,000 matches from 52 soccer leagues and 35 countries, was made available to participants and remains publicly available. As mentioned, this dataset now acts somewhat as a benchmark dataset in this domain despite not containing match features other than goals scored. The omission was likely a deliberate design choice by the challenge organizers, who likely integrated only data that are readily available for most soccer leagues worldwide — including lower-division leagues — so as to maximize the size of the dataset. The papers related to the top-performing models from the 2017 Soccer Prediction Challenge were published in a special issue of the Machine Learning (Springer) Journal. Challenge participants trained models using the 216,000+ match training dataset to predict 206 future match results. The Ranked Probability Score (RPS) [42, 29], which measures how good forecasts are compared to observed outcomes when the forecasts are expressed as probability distributions, was used as the evaluation metric in the competition. Specifically, the participants' models were evaluated based on the RPS averaged across the 206-match challenge test set.

Table 1 Datasets that can potentially be used for ML for soccer match result prediction. The attributes for each dataset can be confirmed by navigating to the corresponding URL.

Dataset Name	Description	URL(s)
European Soccer Database	Contains data on 25,000 matches and 10,000 players in 11 European leagues from 2008 – 2016. Player and team attributes are from the EA Sports FIFA video game. The database also contains team lineups and formations (x,y coordinates), as well as betting odds from up to 10 bookmakers. For over 10,000 matches, the database also contains events, e.g., goal types, possession, corners, crosses, fouls, and cards.	kaggle.com/datasets/hugomathien/soccer
Betting websites	Betting sites provide betting odds and sometimes additional features (but sometimes only short-time periods are available for free).	football-data.co.uk, betfairdatascientists.github.io, and oddsportal.com
engsoccerdata	An R package [34] that provides English and other European league data, along with US MPL and South African league data.	github.com/jalapic/engsoccerdata
Wyscout event data	Pappalardo et al. [90] and Wyscout provided spatiotemporal event data containing 1,941 matches from the top 5 European Soccer leagues, EURO 2016 and the 2018 World Cup.	doi.org/10.6084/m9.figshare.c.4415000.v5
Statsbomb open data	Primarily event log data (with lineup and match metadata).	github.com/statsbomb/open-data
Open International Soccer Database	Contains the match results (season, league, date, home team, away team(Sea Lge Date HT AT HS AS GD WDL) of 216,743 games played between 19/03/2000 and 21/03/2017 from 52 leagues in 35 countries. This dataset was used in the 2017 Soccer Prediction Challenge.	osf.io/kqcye/
soccerdata	A scraper collection implemented in Python that scrapes soccer data from various websites including Club Elo, ESPN, FBref, FiveThirtyEight, Football-Data.co.uk, SoFIFA, and WhoScored.	github.com/probberechts/soccerdata
World Football Elo Ratings	Contains the current Elo ratings of national soccer teams.	eloratings.net
Football Database	Contains the current Elo ratings of club teams from various leagues around the world.	footballdatabase.com
understat.com	Contains expected goals (xG) (however, the way in which these are computed is relatively opaque).	understat.com
football-data.co.uk	Contains, for various leagues, betting odds from multiple providers as well as match statistics for some seasons.	football-data.co.uk
FIFA Index	Contains player and team ratings from the EA Sports FIFA video game	fifaindex.com

2.2.1 2017 Soccer Prediction Challenge: Top Performers

We now cover three of the four papers that achieved the best performance in the competition.[3]

[3] The fourth top-performing paper [116] is not covered in this chapter because the methods used were more statistical in nature: e.g., a Bradley–Terry model, Poisson log-linear hierarchical model, and integrated nested Laplace approximation.

Hubáček, Sourek & Zelezny [59] used relational- and feature-based methods to build models using the Open International Soccer Database. Pi-ratings and PageRank ratings were computed for each of the teams in each match. Both the regression and classification forms of XGBoost [26] were employed as the feature-based method, while boosted relational dependency networks (RDN-Boost) [85] were used as the relational method. The classification with XGBoost on the pi-ratings feature set performed best on both the validation set and the unseen challenge test set, achieving 0.5243 accuracy and an average RPS of 0.2063. Adding more model features, weighting aggregated data according to recency, and including expert guidance by using, e.g., active learning, were mentioned as possible avenues for further work.

Constantinou [28] created a model that combined dynamic ratings with a Hybrid Bayesian Network. The rating system used was a modified version of pi-ratings, which the author had proposed in previous work [30]. The computation of pi-ratings emphasized the match result (win, draw, or loss) to a greater degree than the goal margin, with the aim of dampening the influence of large goal differences. In contrast to the original pi-ratings system, this version also incorporated a team form factor that aims to identify continued over- or under-performance. Four rating features (two for the home team and two for the away team) were used as the inputs to the Hybrid Bayesian Network. Notably, the proposed Hybrid Bayesian Network with modified pi-ratings was able to effectively predict a match between two teams, even when the prediction was based on historical match data that involved neither of the two teams. On the challenge test set, accuracy of 0.5146 and an average RPS of 0.2083 was obtained. Incorporating other key features or expert knowledge, e.g., player transfers, key player availability, international competition participation, management, injuries, attack/defence ratings, and team motivation/psychology, were mentioned as potential opportunities for future improvement.

Berrar et al. [8] applied two ML models to two different feature sets: recency features and rating features. Recency features were computed by averaging feature values over the previous nine matches and were based on four feature groups: attacking strength, defensive strength, home advantage, and opposition strength. A new soccer-specific rating system was proposed, subsequently referred to as "Berrar ratings."[4] XGBoost [26] and k-Nearest-Neighbors (k-NN) were applied to each of the two feature sets, with both models performing better on the rating features compared to the recency feature set. XGBoost applied to the rating features provided the best performance with accuracy of 0.5194 and an average RPS of 0.2054; however, this result was obtained after the challenge had concluded.[5] The authors noted that soccer's low-scoring nature and narrow margins of victory make it challenging to predict based only on goals, while also emphasizing the importance of effective feature engineering and incorporation of domain knowledge.[6] The authors also suggested including match features (e.g., yellow/red cards, fouls, possession, passing and running rates), player-level characteristics (e.g., salaries, ages, and physical

[4] Please refer to subsection 4.1.3 for a description of how Berrar ratings are calculated.

[5] k-NN applied to the rating features achieved 0.5049 accuracy and an average RPS of 0.2149 on the challenge test set, which was the best result achieved during the competition itself.

[6] Incorporation of domain knowledge is discussed in the Appendix to this chapter.

Table 2 Ranked probability scores (RPS) and accuracy results from the 2017 Soccer Prediction Challenge and subsequent studies that have used the Open International Soccer Database

Approach	RPS_{avg}	Accuracy	Paper
CatBoost+pi-ratings	0.1925	0.5582	Razali et al. [96]*
TabNet+pi-ratings	0.1956	0.5582	Razali et al. [97]*
Bookmaker odds	0.2020	0.5194	Bookmakers [100]*
Elo Ordered Logit Model	0.2035	0.5146	Robberechts & Davis [100]*
XGBoost+Berrar ratings	0.2054	0.5194	Berrar et al. [8]*
XGBoost+pi-ratings	0.2063	0.5243	Hubáček et al. [59]
Double Poisson	0.2082	0.4888	Hubáček et al. [60]**
Hybrid Bayesian Network+pi-ratings	0.2083	0.5146	Constantinou [28]
Bradley–Terry, Poisson log-linear hierarchical model, integrated nested Laplace approximation	0.2087	0.5388	Tsokos et al. [116]
Berrar ratings	0.2101	0.4854	Hubáček et al. [61]*
kNN+pi-ratings	0.2149	0.5049	Berrar et al. [8]

Studies that were not part of the 2017 Soccer Prediction Challenge are denoted by *.
The ** denotes that this study used a subset of the Open International Soccer Database, from the 2000/2001 to 2005/2006 seasons.

conditions), and team-level characteristics (e.g., average height, attack running rate) could result in improved model performance.

Notably, the three top-performing studies in the 2017 Soccer Prediction Challenge applied machine learning models — namely, gradient-boosted tree models [8, 59] and Bayesian Networks [28] — to feature sets consisting of soccer-specific rating features.

2.2.2 Post-challenge studies that have used the Open International Soccer Database

Following the conclusion of the 2017 Soccer Prediction Challenge, several other researchers have utilized the Open International Soccer Database. The approaches and results of these studies in terms of RPS and accuracy are summarized, along with the results from the top 4 2017 Soccer Prediction Challenge studies, in Table 2.

Robberechts & Davis [100] used goals- and result-based offensive/defensive models and Elo ratings to predict FIFA World Cup results as well as matches in the Open International Soccer database. Elmiligi & Saad [40] also used the Open International Soccer Database, and proposed a hybrid approach that combined machine learning and statistical models, achieving 0.4660 accuracy and an average RPS of 0.2176. The authors compared the performance when using all matches versus only part of the match history, and when using individual league models versus a model for all leagues. Razali et al. [96] made use of the Open International Soccer Database, comparing the performance of gradient-boosting algorithms including XGBoost, LightGBM [71], and CatBoost [93] on goals- and result-based Elo ratings [65, 100],

as well as pi-ratings. The authors found that applying CatBoost to pi-ratings yielded the best performance, with an average RPS of 0.1925 and accuracy of 0.5582, which outperformed the previous 2017 Soccer Prediction Challenge studies (and other studies that have used the Open International Soccer Database). In another recent study, Razali et al. [97] used a deep learning approach called TabNet [3], a deep neural network designed for tabular data, achieving an average RPS of 0.1956 and accuracy of 0.5582. In a study that only used rating systems and statistical models for results prediction, Hubáček, Šourek & Železný [61] found that Berrar ratings provided the best performance (average RPS: 0.2101, accuracy: 0.4854), followed by Bivariate Poisson, Double Poisson, Double Weibull, and pi-ratings, all of which obtained an average RPS of 0.2103. In earlier work, Hubáček, Šourek & Železný [60] found that on a subset of the Open International Soccer Database (the 2000/2001 to 2005/2006 seasons), the Double Poisson model provided the lowest average RPS of 0.2082 and accuracy of 0.4888 (pi-ratings, however, provided slightly higher accuracy).

Another soccer prediction challenge, using a similar dataset, was held in 2023, however, the results were not available at the time of writing. Please see the appendix of this chapter for a brief description of the 2023 Soccer Prediction Challenge.

3 Models

In this section, we first discuss the need for a clear model objective. Given this model objective, a guide to selecting a set of candidate models, as well as a list of commonly applied traditional machine learning models in this domain[7] is provided. Then, we discuss model interpretability and its greater importance to some groups relative to others. We then discuss gradient-boosted tree models including XGBoost and CatBoost, which have provided some of the strongest performance in this domain of late. Following this, deep learning models, which have shown great promise in a number of application domains, are discussed in the context of soccer match result prediction.

3.1 Model Objective

Models can vary in their objectives, and it is important to establish what the objective of a model is at the outset of a match result prediction project or study. The objective of a model might be purely to achieve high predictive performance, e.g., to compete with expert predictions or in competitions. The model's objective may instead/also be used to place bets on match results. If the predictive model is used for betting, there also needs to be consideration of which matches should be bet on, e.g., using criteria such as the Kelly Index [114]. Finally, a model might be used for performance analysis

[7] By traditional, we mean excluding gradient-boosted tree and deep learning models, which are relatively recent developments in this domain and are described later in this chapter.

purposes by coaches or analysts. Model interpretability is important for this group so that they are able to identify the most relevant features that are of importance to winning so that past performance can be analyzed and team strategy can be adjusted to increase the chance of winning future matches. Model interpretability is discussed further in subsection 3.3.

3.2 Candidate Models

When deciding on a set of candidate ML models for soccer match result prediction, a useful starting point is to thoroughly review the literature and identify models that have performed well in predicting sports match results in general, but also those that have been effective in soccer in particular. The purpose of the model is again relevant when selecting the set of candidate models, as well as who the target audience is since, as mentioned, interpretability is of greater importance to certain groups. On the other hand, if the purpose of the model is purely to achieve the highest performance, black-box models can be sufficient. While some studies have used a single type of model to predict match results, it is more common to conduct a comparative evaluation of the performance of a range of candidate ML models. Furthermore, as well as the 2017 Soccer Prediction Challenge, to confirm model generalizability, some researchers have applied their models to several leagues, e.g., the top 5 European leagues [35, 74, 127], the Greek/Dutch/English leagues [82] and the leagues of 12 different countries [23]. Before recent studies that have used gradient-boosted model trees and deep learning models, the most commonly applied classification ML models for soccer match result prediction included Logistic Regression, Artificial Neural Networks (ANNs), Bayesian Networks, Decision trees, k-NN, Naïve Bayes, Random Forest, and Support Vector Machine (SVM).

3.3 Model Interpretability

Increasing the interpretability (understandability) of match result prediction models has become an area with increasing research activity recently. As mentioned, with greater interpretability, match result prediction models are of greater use to, e.g., coaches and performance analysts, to identify the most relevant features that are within players' control and can be adapted to increase the chance of winning future matches. Models with greater interpretability can also be used by teams and/or management to adapt their tactical decisions, to identify own and opposition team strengths and weaknesses, decide on appropriate formations, and make player selection and transfer decisions [124]. Moustakidis et al. [84] used explainable ML models using SHapley Additive exPlanations (SHAP) [80] to identify team performance indicator metrics that are of greatest relevance to predicting teams' average scoring performance per season. Another study that used the SHAP method is that

of Ren & Susnjak [99]. Yeung, Bunker, & Fujii [124] proposed an interpretable ML model framework for soccer match result prediction that is conceptually similar to GAP Ratings [118], but that predicts match statistics with — rather historical match statistics, which can be cumbersome to engineer — management decision- and player quality-related features. Player quality features were obtained from the EA Sports video game FIFA, a data source that has been used in other studies, e.g., [35, 92]. Random forest feature importance is another appealing approach to interpreting the most relevant model features and has been used in conjunction with a Hybrid Random Forest model by Groll et al. [53]. Of course, some traditional ML models, e.g., decision trees and logistic regression, are also highly interpretable by design and may be useful if they can provide sufficient performance.

3.4 Ensemble Methods

Ensemble methods in ML can be broadly grouped into boosting methods, e.g., gradient-boosted tree models such as XGBoost and CatBoost, and bagging methods, e.g., Random Forest.

As previously mentioned, at least in the absence of match features other than goals scored, the 2017 Soccer Prediction Challenge and subsequent studies using the Open International Soccer Database have highlighted that gradient-boosted tree models, applied to soccer-specific ratings, are currently able to achieve some of the highest performance in this domain (Table 2).

However, on other datasets (some of which include match statistics), Random Forests have been found to be competitive with [5] and even exceed the performance of gradient-boosted tree models [107, 1, 43]. Despite ensemble methods appearing to provide generally better performance than statistical and other traditional ML models, a thorough comparative evaluation of the performance of ensemble methods with deep learning and (deep) neural network models is still required.

Extreme gradient boosting, or XGBoost [26], is one of the most popular gradient-boosted tree methods and has performed well in a variety of ML tasks. The foundation of XGBoost lies in gradient boosting. Gradient boosting, which can be used for both regression and classification, combines several weak learners (e.g., decision trees) into so-called strong learners (gradient-boosted trees). The gradient boosting process sequentially builds simple weak prediction models that each predict the residual error of the preceding model. The weak learners are added to the ensemble and their contributions are determined based on a gradient descent optimization problem, which minimizes an overall loss function representing the difference between the actual results and predictions of the strong learner.

CatBoost [93] has some similarities to XGBoost and is capable of handling both regression and classification tasks by aggregating multiple weak learners. One of the distinctive features of CatBoost is its ability to handle categorical features efficiently, alleviating the need for extensive data preprocessing. CatBoost employs a technique called “ordered target encoding,” which involves encoding a categorical

feature sequentially while excluding the needs of the target feature. This prevents information leakage (where the model accesses the target information of the current observation), ensuring an unbiased encoding process and preventing overfitting. The generation of match outcome probabilities (e.g., for win/draw/loss) that are well-calibrated is a distinct advantage of CatBoost, given its importance in match result forecasting [121].

Another boosted-tree model that has not been widely investigated, but that may have potential, for soccer match result prediction is the Alternating Decision Tree (ADTree) model [46]. As opposed to XGBoost and CatBoost, ADTree uses AdaBoost [47] rather than gradient boosting, and maintains an interpretable decision tree structure as the final model (see the Appendix for more detail).

3.5 Deep Learning Models

Deep learning models have been effective in various domains including computer vision, trajectory analysis, and event prediction in sports. Nevertheless, there are still only a relatively small number of published articles related to deep learning for soccer match result prediction.

Danisik, Lacko, & Farkas [35] compared the performance of a Long Short-Term Memory (LSTM) model [56] with classification, numeric prediction, dense approaches, and also against average random guess, bookmaker prediction, and home team win baselines in predicting English Premier League matches. Rahman [94] used deep neural networks and ANNs to predict match results from the 2018 FIFA World Cup. Jain, Tiwari & Sardar [67] used Recurrent Neural Networks and LSTM networks for predicting English Premier League match results, engineering several features related to winning and losing streaks, points, and goal differences. Malamatinos, Vrochidou, & Papakostas [82] used k-NN, LogitBoost, SVM, Random Forest, and CatBoost — as well as convolutional neural networks and transfer learning with encoded tabular data converted to image models — to predict Greek Super League match results, with the best performing model also applied to the English Premier League and the Dutch Eredivisie. The best-performing model was found to be CatBoost, notably outperforming the deep learning convolutional neural network model. Given the time-series nature of soccer match result data, Joseph [69] considered time series-based approaches to predict English Premier League match results, including LSTM and Bayesian methods. As mentioned in subsection 2.2.2, Razali et al. [97] recently achieved strong performance (Table 2) when applying TabNet [3] — a deep neural network model for tabular data — to the Open International Soccer Database.

Given the relatively small number of studies related to deep learning in soccer result prediction, there is the potential for greater future exploration into the potential of these models, e.g., to investigate whether deep learning models are able to outperform boosted tree and other ensemble models.

4 Features

This section describes some of the different types of features that can be used in soccer match result prediction models.

Features in sports match result data can generally be divided into different feature subsets. For instance, in basketball, Miljković et al. [83] categorized features into match-related and standings features. Tax and Joustra [113] compared the performance of a combined feature set consisting of betting odds and features derived from publicly available data to a feature set consisting only of betting odds. Hucaljuk and Rakipović [62] compared the performance of a separate expert-selected feature set against their own selected feature set. Feature selection algorithms can be applied to the entire feature set as well as subsets (e.g., feature selection algorithm-selected vs. human-selected features, betting odds included vs. betting odds excluded, rating features, match features, external features, etc.) of the original feature set, and the performance of the candidate models on each feature set can be compared.

In the remainder of this section, first, rating features are discussed, including general-purpose rating systems that are used across a range of sports, e.g., Elo ratings. Domain-specific (soccer-specific) rating systems including pi-ratings, Berrar ratings, and the recently proposed GAP Ratings are then discussed, as are betting odds (which can also be considered a rating for which the calculation is opaque). Then, match, player, and team statistics are examined, as are external features that do not relate to events occurring within matches. Feature selection methods are also briefly discussed.

4.1 Ratings

The general idea of rating systems is to assign some initial rating for each team and update the ratings over time based on actual match results. Some soccer-specific ratings also consider the goal margin, match venue, and offensive/defensive strengths. There are several use cases for ratings, the first and most obvious of which is to rate/rank teams. In the context of match results prediction, rating systems can, of course, be used as a predictive model by simply predicting the team with the higher rating as the winner. However, using ratings as model features in machine learning models appears to provide better performance (Table 2), and for this reason, ratings are described here in the features section of this chapter rather than in the models section.

4.1.1 Elo Ratings

The Elo ratings [41] system was originally used in chess but has since been applied in a wide range of sports including soccer. Extensions of Elo have been proposed, e.g., to account for goal margin [65] and home advantage [104].

In the Elo rating system, each team generally starts with a rating of 1500. The system can be decomposed into two steps: the estimation step (E-step) and the update step (U-step). The E-step involves estimating the probability that a team will win a particular match, while the U-step involves the update of the team's rating following a particular match result.

E-step. The probability that team i beats team j is estimated by:

$$p_{i,j}(t) = \frac{1}{1 + 10^{\frac{-(R_i(t) - R_j(t))}{400}}} \tag{1}$$

Analogously, the probability that team j beats team i is:[8]

$$p_{j,i}(t) = \frac{1}{1 + 10^{\frac{-(R_j(t) - R_i(t))}{400}}} \tag{2}$$

Since these are probabilities, they add to one, i.e., $p_{i,j}(t) + p_{j,i}(t) = 1$.

U-step. The Elo ratings are updated, e.g., for team i, according to:

$$R_i(t+1) = R_i(t) + K[A_i(t) - p_{i,j}(t)]$$

Team j's rating is updated in a similar manner:

$$R_j(t+1) = R_j(t) + K[A_j(t) - p_{j,i}(t)]$$

where $A_i(t)$ takes the value of 1, 0.5, or 0 for a win, draw, and loss, respectively.

K-factor. The K term in the above U-step expressions is known as the K factor and is often given a fixed value (e.g., 32 or 20). If K is too large, the Elo rating will vary excessively from match to match; however, if K is too small, the rating of a team will not change rapidly enough when it improves its performance [65].

There are different approaches to selecting the value of K: it can be a fixed value, a tunable parameter, or a function. Hvattum & Arntzen [65] proposed a goal-based Elo rating system that allows K to depend on the goal difference. In particular, $K = K_0(1 + \delta)^{\lambda}$, where δ is the absolute goal margin, and K_0 and λ are parameters.

Other approaches to selecting the K-factor are described online, e.g., on eloratings.net and footballdatabase.com (see the Appendix for further details).

Elo Ratings with Home Advantage. Ryall & Bedford [104] extended Elo ratings to incorporate home advantage (albeit in Australian Rules Football, but it can equally be applied in soccer) by simply adding a home advantage term to the rating difference such that the estimated probability of team i beating team j from Equation 1 becomes

$$p_{i,j}(t) = \frac{1}{1 + 10^{\frac{-(R_i(t) - R_j(t)) + H_{i,j}}{400}}} \tag{3}$$

[8] The value of 400 was specified by Elo [41] to be twice the standard deviation of chess player ratings. The values of 10 and 400 merely serve to set a scale for the ratings and are generally used when the initial ratings are set to 1500 [65].

where $H_{i,j}$ is the magnitude of home advantage that team i has over j. Note that in this setup, $H_{i,j}$ is a tunable parameter that does not vary by match, t.

Connection between Elo and Bradley-Terry. As mentioned earlier, the Bradley-Terry paired comparisons model [11], which estimates the win probability in a match using the strengths/abilities of each team, is the foundation of Elo ratings. In the Bradley-Terry model, the probability of team i beating team j is given by:

$$p(i \text{ beats } j) = \frac{s_i}{s_i + s_j}$$

where s_i and s_j denote the strength (ability) of teams i and j, respectively. The strength parameters are commonly estimated using maximum likelihood, e.g., based on observed win counts between the two teams. The Elo rating of a team can actually be expressed directly as a function of its Bradley-Terry strength [33]. In particular, the Elo rating of team i can be expressed as $R_i = 400log_{10}(s_i)$. Rearranging this expression enables one to calculate the Bradley-Terry strength of team i from its Elo rating: $s_i = 10^{\frac{R_i}{400}}$.

The Bradley-Terry model has been shown to be outperformed in soccer match result prediction by Poisson/Weibull-based models [61]. Furthermore, when using the ratings themselves for prediction, pi-ratings have been found to generally outperform Elo ratings.

4.1.2 Pi-ratings

In the pi-ratings system [30], each team is assigned two ratings corresponding to their home and away strengths, with their overall rating the average of their home and away ratings:

$$R_i = \frac{R_{i,H} + R_{i,A}}{2}$$

Then, for each match, the expected goal difference is calculated based on the home rating of the home team and the away rating of the away team. The actual match outcome is compared with the expected score and if a team performs better than expected, its rating is increased based on the difference between actual and expected outcomes, as well as the learning rates (which are model parameters). The home and away ratings of both teams are updated after a match but using separate learning rates. Each team starts with an initial pi-rating of 0, and the pi-rating represents the rating of the average team relative to other teams. Given a match between teams α and β, the ratings are updated according to:

$$\hat{R}_{\alpha,H} = R + \lambda\psi_H(e)$$

$$\hat{R}_{\alpha,A} = R_{\alpha,A} + \gamma(\hat{R}_{\alpha,H} - R_{\alpha,H})$$

$$\hat{R}_{\beta,H} = R_{\beta,H} + \lambda\psi_A(e)$$

$$\hat{R}_{\alpha,H} = R_{\alpha,H} + \gamma(\hat{R}_{\alpha,A} - R_{\alpha,A})$$

where e is the error between the expected and actual goal difference. The expected goal difference for a team against the average opponent at ground G, which is either H (home) or A (away), is calculated as:

$$\hat{g}_{DG} = b^{\frac{|R_{iG}|}{c}} - 1$$

where $b = 10$ and $\hat{g}_{DG}$ is the expected goal difference for team i against the average opponent when playing at ground G (either H or A). Team ratings can potentially be negative, in which case the expected outcome is $-\hat{g}_{DG}$. The error, e, between the expected and actual goal difference is then given by:

$$e = |g_D - \hat{g}_D|$$

where $g_D = g_{DH}$ - g_{DA} and $\hat{g}_D = \hat{g}_{DH}$ - $\hat{g}_{DA}$ To dampen the influence of large goal margins when updating ratings, a reward/penalty function is specified to be a function of the goal difference error:

$$\psi(e) = c \times log_{10}(1 + e)$$

where $c = 3$. The probability distribution over potential match outcomes is computed using an ordered logit model.

4.1.3 Berrar Ratings

In the Berrar ratings system [8], a logistic function is used to predict the expected number of goals scored using the teams' offensive and defensive strengths. In particular, the expected goals scored by the home and away teams are respectively given by:

$$\hat{G}_H = \frac{\alpha_H}{1 + exp(-\beta_H(o_H - d_A) - \gamma_H}$$

and

$$\hat{G}_A = \frac{\alpha_A}{1 + exp(-\beta_H(o_A - d_H) - \gamma_A}$$

where o_H and d_H and o_A and d_A denote the offensive and defensive ratings of the home and away teams, and α_H and α_A represent the maximum possible expected goals that can be scored in a match by the home and away teams. The β parameters determine how steep the logistic function is, and γ determines the threshold/bias.

The ratings for the home team are updated according to:

$$o_H = o_H + \omega_{o_H}(G_H - \hat{G}_H)$$

$$d_H = d_H + \omega_{d_h}(G_A - \hat{G}_A)$$

And similarly, for the away team:

$$o_A = o_A + \omega_{o_A}(G_A - \hat{G}_A)$$

$$d_A = d_A + \omega_{d_A}(G_H - \hat{G}_H)$$

4.1.4 GAP Ratings

Since soccer is a low-scoring sport and goals are rare, more recently, Generalized Attacking Performance (GAP) Ratings [118] were proposed, which build on pi-ratings but can also instead predict non-rare match statistics (e.g., shot attempts), not only goals scored. By predicting non-rare match statistics and excluding some of the early and latter rounds of seasons, GAP ratings were shown to be more informative (in terms of the Akaike Information Criterion) for soccer match result forecasting [120].

GAP ratings [118, 120] are inspired by pi-ratings [30] and are adjusted such that they can be applied to predict match statistics in advance. The GAP rating requires match statistics as input and can be used to predict goals, shots on and off target, and corners for both teams. The inputs, S_H and S_A, denote the match statistics for the home team (H) and away team (A), respectively. There are four ratings for each team i: H_i^a, H_i^d, A_i^a, A_i^d, which denote team i's home attack, home defence, away attack, and away defence ratings, respectively. Following a match between home team i and away team j, the GAP ratings for home team i are updated as follows:

$$H_i^a = max(H_i^a + \lambda\phi_1(S_H - \frac{H_i^a + A_j^d}{2}), 0)$$
$$A_i^a = max(A_i^a + \lambda(1-\phi_1)(S_H - \frac{H_i^a + A_j^d}{2}), 0)$$
$$H_i^d = max(H_i^d + \lambda\phi_1(S_A - \frac{A_j^a + H_i^d}{2}), 0)$$
$$A_i^d = max(A_i^d + \lambda(1-\phi_1)(S_A - \frac{A_j^a + H_i^a}{2}), 0)$$

The GAP ratings for away team j are updated as follows:

$$A_j^a = max(A_j^a + \lambda\phi_2(S_A - \frac{A_j^a + H_i^d}{2}), 0)$$
$$H_j^a = max(H_j^a + \lambda(1-\phi_2)(S_A - \frac{A_j^a + H_i^d}{2}), 0)$$
$$A_j^d = max(A_j^d + \lambda\phi_2(S_H - \frac{H_i^a + A_j^d}{2}), 0)$$
$$H_j^d = max(H_j^d + \lambda(1-\phi_2)(S_H - \frac{H_i^a + A_j^d}{2}), 0)$$

where $\lambda > 0$ is a parameter that weights the effect of the latest game on the rating, and $\phi_1 \in [0, 1]$ and $\phi_2 \in [0, 1]$ weights the effect of a home game on the away rating and an away game on the home rating, respectively.

Using goals as an example, H_i^a indicates the number of goals home team i will score and A_j^d indicates the number of goals the away team j will concede. The average of H_i^a and A_j^d represents the expected goals home team i should score and away team j should concede. Ratings are updated using the weights and the differences between actual and expected results. The GAP rating can be interpreted as the performance of a team compared to an average team in the league. Better teams will have higher attacking and lower defensive ratings. The estimates of S_H and S_A, which are denoted by $\widehat{S_H}$ and $\widehat{S_A}$, respectively, are given by:

$$\widehat{S_H} = \frac{H_i^a + A_j^d}{2}$$

$$\widehat{S_A} = \frac{A_j^a + H_i^d}{2}$$

To select the λ, ϕ_1, and ϕ_2 parameters, the least-squares method is applied. The cost function is given by:

$$Cost = \sum_{i=1}^{N} |S_H - \widehat{S_H}| + |S_A - \widehat{S_A}|$$

where N denotes the number of matches. All GAP ratings are initialized to zero. Parameter selection is repeated at the start of each season using all data. However, the first year of data is only used for parameter selection, and the first six games and final six games played by the home team in each season are ignored due to various factors that could increase the unpredictability of these matches. Also, the relegated teams' ratings are calculated as the average of the promoted teams' ratings, and similarly, the promoted teams' ratings are calculated as the average of the relegated teams' ratings.

4.1.5 Betting Odds

As mentioned, betting odds can be used as a baseline with which to model results (subsection 4.1.5), but they can also be considered a type of rating [122]. Therefore, like other ratings, odds can potentially be used as a model feature. Indeed, betting odds have been used as model features in many studies [106, 86, 113, 44], sometimes as the sole model feature. Since bookmakers establish betting odds, unlike the rating systems described above, the way in which odds are determined is opaque. However, it is reasonable to assume that bookmaker companies use models supplemented with observed betting volumes on each outcome to capture the "wisdom of crowds" [111, 17] in order to set appropriate odds for matches. Betting odds also differ from other ratings that generally only incorporate historical match result information, since

odds also account for other factors such as market sentiment, player availability, and expert knowledge.

Deciding whether betting odds should be included as a model feature depends on the objective of the model. For instance, if a model is being used for betting purposes in an attempt to "beat the house," betting odds should not be included as a model feature [58]. As a baseline, betting odds have proven difficult to beat. For example, despite using a state-of-the-art gradient boosting model along with a sophisticated engineered set of features, Baboota & Kaur [5] were unable to outperform predictions based on bookmaker odds. Bookmaker predictions also outperformed the top-performing submissions from the 2017 Soccer Prediction Challenge (odds outperformed all competing models in [100], and this result is also shown in Table 2, along with the next-best result from that study, an Elo ordered logit model). On the other hand, if the objective is purely maximizing predictive performance, given the information inherent in betting markets, including betting odds model features may make sense.

Some studies have attempted to exploit bookmaker odds while using machine learning models, e.g., through arbitrage [74] and modern portfolio theory methods [58], including Talattinis et al. [112], who utilized the Sharpe Ratio. Many betting sites provide betting odds data for future and historical matches (Table 1). When sourcing betting odds from such websites, to create a model feature, it is often useful to convert the raw decimal odds into probabilities by taking their reciprocal:

$$\text{Probability} = \frac{1}{\text{Decimal Odds}} \tag{4}$$

The sum of the odds generally exceeds one due to the presence of a built-in bookmaker margin, which is sometimes referred to as the "over-round" [45, 125]. To ensure that the odds add to one and obtain the normalized odds-implied win probabilities, one simply divides the original odds-implied probabilities by the normalization factor, which is the sum of the probabilities obtained for each match outcome from equation 1.4. That is,

$$\text{Probability Normalized } = P_{norm} = \frac{\text{Probability}}{\text{Normalization Factor}}$$

For example, suppose that a match has only two possible outcomes and that the decimal bookmaker odds are such that team A is paying 1.34 to win and team B is paying 3.02. Taking the reciprocal as per equation 1.4, we obtain 0.746 and 0.331 for teams A and B, respectively. Then, 1.077 is the normalization factor ($0.746+0.331 = 1.077$). To obtain the odds-implied win probabilities, we have $P_{norm}(\text{team A win}) = 0.746/1.077 = 0.693$ and $P_{norm}(\text{team B win}) = 0.331/1.077 = 0.307$, so now $P_{norm}(\text{team A win}) + P_{norm}(\text{team B win}) = 0.693 + 0.307 = 1$.

Of course, betting on the match outcome is not the only option for bettors: it is possible to bet on a specific event occurring within a match, who will score the first goal, and so on. Betting on the over-under — whether a team scores over or under 2.5 goals in a match — is one such option. Wheatcroft [118] found that in the over/under

2.5 goals market, shots made and corners were better for probabilistic forecasting than actual goals.

4.2 Match Features

Simple Match Statistics & Performance Indicator Metrics: As has been mentioned in previous sections, match features (also known as in-play, in-match, or in-game features) are features that relate to events that occur within matches. Match features are often performance indicator metrics [63], as they are referred to in the field of sports performance analysis, a sub-discipline of sports science. For instance, raw statistics such as passes, tackles, shot attempts, etc., can be aggregated into performance indicator metrics at the player, offensive/defensive unit, or team levels. The survey paper of Bunker & Susnjak [18] highlighted the potential benefits of greater future collaboration between researchers from sports performance analysis and machine learning for engineering relevant match features for machine learning models.

Of course, match features that are performance indicators are not known in their entirety until the match has concluded. Therefore, for match result forecasting, preprocessing is required, by aggregating (e.g., averaging) match features across a certain number of historical matches. There is no set number of matches with which to aggregate/average over since it will depend on the characteristics of the dataset. For example, Buursma [22] found that aggregating features over the past 20 matches provided the best performance, while Berrar et al. [8] found that using the past nine matches provided the best performance in building a set of what the authors referred to as "recency features" for the 2017 Soccer Prediction Challenge. Simple aggregation (e.g., averaging or summing) of features over a certain number of past matches does not, however, account for the fact that recent games are of greater relevance than matches that were played a long time ago. Exponential time weighting [37, 78, 60], which is used in the weighted likelihood function that is maximized in Double Poisson models, could be used as an alternative to simple aggregation over historical matches.

Matches can be summarized in the form of simple statistics, some of which are performance indicators and some are not. Goals are the most commonly used match statistic since goals are also used to derive the match result, and are therefore used in essentially every study on match result forecasting. Shots on target, shots off target, and corners are other commonly used match statistics. Other match statistics in soccer that can also be considered performance indicators include blocks, clearances, crosses, dribbles, interceptions, passes, possessions, saves, and tackles. Discipline-related match statistics include yellow and red cards, fouls, and offsides. As mentioned, match statistics prediction is an important part of the GAP Ratings method [118], which focuses on predicting non-rare events such as shot attempts (since actual goals are rare).

On-the-Ball Event data: Different vendors and data providers, e.g., StatsBomb and Stats Perform/Opta, have varying data formats and definitions that are used to describe and annotate events that occur within soccer matches. Decroos et al. [36] proposed a Soccer Player Action Description Language (SPADL) to standardize the event data format across vendors. The authors also proposed a method called Valuing Actions by Estimating Probabilities (VAEP), which uses event log data that has been converted into SPADL format, to value on-the-ball actions. VAEP can also be used to compute player ratings by aggregating each of their on-the-ball actions. Event data commonly includes the start and end times and locations of events, the player who performed the action and the team to which they belong, the action type (e.g., pass, cross, throw in, etc.), the body part used in the action, and whether the action was successful or unsuccessful. While event data focuses event on on-the-ball rather than off-ball events, supplementing on-the-ball event features with tracking data-derived features can address this.[9] Event data can provide more in-depth information than simple match statistics and has the potential to be further explored for engineering informative features for soccer match result prediction ML models.

Spatiotemporal tracking data: Spatiotemporal tracking data is generally sourced from optical systems or wearable (GPS) tracking devices that record the locations of players and the ball. Another way in which spatiotemporal data can be obtained is from computer vision, machine learning, or deep learning methods applied to match video footage. As mentioned, it is valuable for event data and spatiotemporal data to be combined to be able to perform more holistic performance analyses, however, this is challenging – both in terms of cost but also technically — because it generally requires collaboration between researchers from disparate disciplines, i.e., computer/data scientists and sports scientists [49]. Furthermore, professional teams are generally unwilling to share data publicly, and obtaining it from professional data providers comes at significant cost. Rein & Memmert [98] also highlighted the value of greater collaboration between sports scientists and computer scientists in extracting tactical insights from spatiotemporal tracking data in soccer, which has traditionally been used for monitoring physiological demands. Toda et al. [115] proposed a method called Valuing Defense by Estimating Probabilities (VDEP), an adjustment of VAEP in which the defensive performance of a team is evaluated using on-the-ball event and tracking data. The generated metrics focus on actions that lead to penetration of the penalty area or a shot attempt (rather than a goal, which is rarer). The study showed that this resulted in higher predictability than VAEP. Deriving features from spatiotemporal tracking data also appears to have the potential for further research in match results prediction ML models in soccer. As mentioned, spatiotemporal tracking data has traditionally been used to assess physiological demands on athletes, and Tümer et al. [117] aimed to predict Turkish Super League rankings by applying three ML models – ANN, radial basis function, and linear regression — to physical and technical features (e.g., related to physical demands).

[9] F. Goes (2021), The power of combining tracking and event data, https://www.scisports.com/the-power-of-combining-tracking-and-event-data/

Expected Goals: Expected Goals (xG) refers to the expected number of goals a team should score in a game. For every shot a team takes in a match, the probability of a goal can be calculated using xG models, of which there are many types. Then, by aggregating all of these probabilities, one can obtain the xG for the match. To build an xG model, shot-related features are considered. For example, according to the Bundesliga,[10] the distance and angle to the goal, whether it is a pass or a header, and the type of pass are used. For the algorithm, any type of regression model, e.g., logistic regression or neural network, can be utilized. The target variable takes the value of 1 for a goal and 0 for no goal. Wheatcroft [120] showed that using goals could not provide statistically significant information. Replacing goals with expected goals (xG) as the target variable is a commonly suggested approach for further research because expected goals are not as rare.

4.3 Player Statistics

Player characteristics and ratings. The incorporation of player-level attributes or ratings has been a popular line of enquiry in recent times. For example, Stübinger, Mangold, & Knoll [108] incorporated player and match attributes in an ensemble learning approach in a simulation on the top 5 European leagues. Chen [25] used FIFA video game player ratings to predict Spanish La Liga soccer results using three common ML models: ANN, Random Forest, and SVM. Another study that utilized FIFA player ratings is that of Danisik, Lacko, & Farkas [35], whose deep learning approach using LSTM has been discussed previously in this chapter. Arntzen & Hvattum [4] compared the performance of an ordered logistic regression model and a competing risk model when applied to adapted plus-minus player ratings — which compare the performance of a team, e.g., in terms of goals, when a particular player is playing versus when they are not [64]) — with when the two models were applied to team Elo ratings. The authors found no difference in performance between the ordered logit and competing risk models for match result prediction. Furthermore, they found that team ratings and player ratings had similar performance, but using a combined feature set with both player ratings and team ratings provided better performance than when utilizing only player ratings or only team ratings. Team and player ratings from EA Sports' FIFA are available via websites such as fifaindex.com (Table 1). Other video games such as Football Manager (Sports Interactive/Sega) also contain player ratings that are arguably better and could potentially be used in future studies.

Player Form: Otting and Groll [88] applied Hidden Markov models to investigate the existence of the "Hot Shoe" effect (known as the "Hot Hand" effect in basketball), and provided evidence that such an effect does exist. If such an effect does exist, it is

[10] Bundesliga (2019), xG stats explained: the science behind Sportec Solutions' Expected goals model, https://www.bundesliga.com/en/bundesliga/news/expected-goals-xg-model-what-is-it-and-why-is-it-useful-sportec-solutions-3177

reasonable to assume that players in the "hot" state are more likely to perform better than those in the "cold" state, which could provide more information for prediction.

4.4 Team Statistics

Team Ratings. In subsection 4.1, team ratings based on historical match results in terms of goals scored were discussed. However, there are other ways to obtain or compute team ratings. The first approach is to obtain team ratings from video games such as FIFA or Football Manager and directly use these as features in a predictive model [5, 124]. Aside from video games, another potential source of team ratings is the Union of European Football Associations (UEFA), which provides team ratings based on results in UEFA club competitions over the past five years [109]. Player ratings can be obtained from video games such as FIFA and then aggregated to engineer role- or team-level ratings [24]. Carpita, Ciavolino and Pasca [24] used 33 player-related features from the FIFA video game to construct seven player-level performance indicators, which were then combined into performance indicators based on role (forward, midfielder, defender and goalkeeper). A binomial logistic regression model was then applied to differences between the role-based performance indicators of opposing teams to analyze the degree to which these indicators affect the probability of winning. Alternatively, player ratings can be computed using VAEP or plus-minus ratings using event log data or historical match result data, respectively, and then aggregated into a team rating.

Some studies have incorporated or compared both player and team ratings. For instance, Pipatchatchawal & Phimoltares [92] incorporated both player and team FIFA video game ratings, while Arntzen & Hvattum [4], as previously mentioned, compared the use of team Elo ratings and player plus-minus ratings.

Streak: Streak is a similar concept to the hot shoe phenomenon; however, it generally refers to match result patterns rather than individual outcomes such as successful shots on goal. Baboota & Kaur [5] computed the team streak as of a team's jth game as:

$$Streak(j) = \sum_{i=j-k}^{j-1} Score_i/3k$$

where $Score_i$ indicates the goals scored in game i, $Score_i \in \{0, 1, 3\}$ and k is a tunable hyperparameter indicating how many historical games are considered. To account for the fact that more recent matches are of greater relevance, the authors also engineered a time-weighted streak feature, which for the jth game, is given by:

$$Weighted\ Streak(j) = \sum_{i=j-k}^{j-1} 2 * \frac{i - (j-k-1)Score_i}{3k(k+1)}$$

The experimental results of the study showed that $k = 6$ provided the best performance for the various models.

Form: The streak only considers a single team; thus, to consider the opponent teams, Baboota & Kaur [5] also engineered a form feature. The form of each team is set to an initial value of one and is updated after each game. The forms of teams A and B, for the jth game between these two teams, if team A wins, are given by:

$$Form_j^A = Form_{j-1}^A + \alpha Form_{j-1}^B$$
$$Form_j^B = Form_{j-1}^B - \alpha Form_{j-1}^A$$

If the match is a draw, the forms are given by:

$$Form_j^A = Form_{j-1}^A - \alpha(Form_{j-1}^A - Form_{j-1}^B)$$
$$Form_j^B = Form_{j-1}^B - \alpha(Form_{j-1}^B - Form_{j-1}^A)$$

The authors found that the hyperparameter $\alpha = 0.33$ provided the best performance.

Inter-player Chemistry: Bransen and Van Haaren [12] suggested that the chemistry between players within the same team is more important than the past performance of a player in another team. Two offensive and defensive chemistry metrics were proposed based on VAEP, which considered the interaction between two players, to identify the squad with the best chemistry. Moreover, two CatBoost models were applied to player statistics features for a pair of players, with the two chemistry metrics as the respective target variables, in order to predict the chemistry of a specific player with players in other teams (which is potentially useful for scouting or transfers).

Passing Networks: Some studies have used social or passing network analysis to predict match results in soccer. Using passing distribution data, Cho, Yoon and Lee [27] used social network analysis to construct actual and predicted network indicators that reflect team performance, in conjunction with gradient boosting for match result prediction. Ievoli, Palazzo, & Ragozini [66] considered whether passing networks have a significant effect on match results, applying four ML models to passing network-derived indicators and an on-field feature set to predict 2016–2017 UEFA Champions League group stage results. The authors found that some network-derived variables were connected to the level of offensive actions, and could improve the explanatory power of models beyond the models that included only on-field features.

4.5 External Features

External features are external to the match itself, i.e., these features are not derived from on-field events. These may include travel, player availability, match venue, match officials, weather, and so on. One study that incorporated weather as a model feature is that of Palinggi [89], who applied an SVM model to weather- and match-related features and achieved around 50% accuracy. Other potential external model

features include the average age of a team's players and coach [52], the number of players who are international representatives [70], as well as club market values, transfer budgets, and operational costs [109]. Unlike match features, external features are generally already known prior to a match and can, therefore, often be used directly as model features, without the type of preprocessing that is required for match features through aggregation over historical matches. In terms of match venue having an effect, the home advantage phenomenon is a well-known effect in which the home team tends to outperform the away team, on average (the COVID-19 pandemic period, where some matches were played in front of empty stadiums, being a possible exception [77]).

Another potential source of external feature data is social media. Wunderlich & Memmert [123] applied sentiment analysis methods to tweets posted on Twitter during over 400 English Premier League matches, as well as those posted in the periods shortly before and after goals, to extract information (e.g., word clouds and other relevant features) to predict the total number of goals scored in a match and perform in-play forecasting. Their results suggested that in-play goal forecasting is very challenging, in-play information did not improve predictive accuracy and its predictive value is small in comparison to pre-match information, and that — perhaps due to overly high pre-match expectations from fans — fan sentiment on Twitter decreases during a match. Kinalioğlu & Kuş [72] proposed a hybrid clustering and classification method and applied it to data from 6,396 European league matches including — as well as team and player statistics — fan opinions from social media. Another study that used the Twitter posts of fans to predict soccer match outcomes is that of Kampakis & Adamides [70].

4.6 Feature Selection Methods

Feature selection techniques can be broadly grouped into three categories: filter, wrapper, and embedded methods [126].

The Random Forest algorithm [13] has embedded tree-based feature selection, in which feature importance is evaluated during the training process [79]. This embedded feature selection means that despite being a bagging method, Random Forest is a model with potentially high interpretability. In the sklearn package in Python, the SelectFromModel object can be used in conjunction with the RandomForestClassifier to automatically select features.

Filter feature selection methods generally rank features based on, e.g., the chi-squared, information gain (ratio), or correlation between each feature and the target variable. A disadvantage of using filter methods is that a cut-off needs to be arbitrarily selected to determine how many features to include in a model.

Other feature selection methods, e.g., CFS subset [55] and Relief/ReliefF subset feature selection methods [73, 75], consider the correlation or interaction among features as well as those between each feature and the target variable. Sequential forward selection is another possible approach to selecting features (e.g., [82]).

As previously mentioned, the performance of models applied to features selected by feature selection techniques can be compared with human expert-selected features (e.g., [62]).

5 Evaluation Methods

Selecting an appropriate evaluation metric is important when evaluating machine learning models. It is also necessary to clearly define the target variable that is to be predicted.

In sports with only two possible outcomes, assuming the league is competitive, class imbalance is generally not an issue since home team wins should only slightly outweigh away team wins due to the home advantage effect. In this case, classification accuracy can be an appropriate evaluation metric.

In soccer, which has three possible outcomes (with a draw much less common than the other two outcomes), accuracy has continued to be commonly used to evaluate classification ML models. However, more recently, scoring rules such as the Ranked Probability Score (RPS) have become prevalent as evaluation metrics, in part due to its use in the 2017 Soccer Prediction Challenge.

5.1 Problem Setup & Target Variable Definition

Soccer match result prediction problems are commonly formulated as a three-class classification problem with a target variable defined with three discrete values (e.g., win/draw/loss),[11] as a numeric prediction problem predicting a goal margin target variable,[12] or by predicting the number of goals scored by each team and thus the match result. Since soccer is a low-scoring sport, numeric prediction of the goal difference is challenging compared to sports with match result margins that are larger in magnitude (e.g., basketball or rugby). In machine learning for soccer result prediction, classification is the most common approach, while statistical models commonly attempt to model and predict the number of goals scored by each team. However, given the 2023 Soccer Prediction Challenge (described further in the Appendix) involved —- as one of the tasks — predicting the number of goals scored by each team, more future studies may begin to consider this problem setup.

Other studies have sought to predict other target variables. For example, predicting the "over-under": whether more than 2.5 goals are scored in total in a match [89],

[11] In classification, some researchers, e.g., [72], excluded draws or merged draws into another class so as to create a binary problem, which naturally gives better performance compared to three-class classification models given the difficulty of predicting draws.

[12] Another approach proposed is to convert win/draw/loss to 1/0.5/0 and use a numeric prediction model. This was an approach followed by [35], who used LSTM regression and found that it outperformed the LSTM classifier.

or forecasting match statistics such as shot attempts or shots on target [118, 124]. Using a Bayesian approach for prediction, Robberechts, Van Haaren, & Davis [101] developed an in-game win probability model. Wunderlich & Memmert [123] considered both in-play forecasting and predicting the total goals scored in a match. Predicting whether both teams will score during a match has also been considered (e.g., by [32]). Others have attempted to predict the final team rankings in a league [117], or whether a team earns a competition point [43]. The per-season average scoring performance of teams has also been considered as a target variable to be predicted [84].

5.2 Baselines

Since datasets in this domain often cover different leagues and seasons, and contain different variables from which model features can be derived, it is generally difficult to compare results across studies unless the studies being compared utilised a common dataset.

Baselines — including simple rule-based benchmarks, betting odds-derived predictions, and expert predictions — are another way in which researchers can evaluate model performance:

- **Betting Odds:** Betting odds are somewhat unique in that they can be included as a model feature but can also act as a benchmark with which to evaluate model performance. It is straightforward to convert decimal odds into outcome probabilities by taking their reciprocal and normalizing such that the probabilities add up to one (subsection 4.1.5). One can then predict the match outcome by simply predicting the outcome that has the highest probability. The odds of several bookmakers can also be combined to create a bookmaker consensus model [76]. Alternatively, provided the model is not being used for betting strategies to "beat the house," these probabilities can potentially be used as model features. The outcome with the highest probability will, of course, almost always be a non-draw outcome.
- **Simple rule-based benchmarks:** A rule that always predicts the majority class (also known as the Zero Rule algorithm) can be used as a baseline. The Zero Rule algorithm (ZeroR), which generally predicts a home team win because of the home team advantage phenomenon, is available in machine learning tools such as WEKA, as are other simple rule-based algorithms such as OneR [57], which can also be used as a benchmark. Random guesses are easily implementable and are also often used as a benchmark, e.g., [35].
- **Expert predictions:** Comparing model predictions with expert predictions, e.g., those of media or former players, is another approach to benchmarking; however,

access to experts can be a challenge.[13] Butler, D., Butler, R., & Eakins [21] compared the accuracy of experts and laypeople over three English Premier League seasons and found that former professional soccer players in particular were superior in predicting match results.

5.3 Scoring Rules

Machine learning models often generate probabilities associated with each class label, and the performance of models — based on how closely these probabilities match the actual (true) class labels — can then be evaluated. Some machine learning models require that match instances are labeled with target variables encoded with a vector with three elements, e.g., in soccer, (1,0,0), (0,1,0) and (0,0,1) can denote win, draw, and loss outcomes, respectively. By representing labels as vectors in this manner, scoring rules such as the Ranked Probability or Brier scores can be utilized, in which a value of 0 indicates a perfect prediction and 1 represents a completely incorrect prediction (thus, lower scores are preferred).

5.3.1 Accuracy

Classification accuracy is a widely used evaluation metric and has also been used in many studies in ML for soccer match result prediction. Accuracy is computed by taking the total number of instances (matches) correctly classified by the model (where the predicted value equals the true value) and dividing this by the total number of instances, i.e.,

$$Accuracy = \frac{C}{N}$$

where C denotes the number of instances correctly classified by the model and N is the total number of instances.

5.3.2 Brier Score (BS)

The original Brier Score [14] can be applied for multi-class prediction. The lower the BS, the smaller the prediction error. The original Brier Score is given by:

$$\text{BS Original} = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{R}(\widehat{y_{in}} - y_{in})^2$$

[13] Rather than getting experts to provide their match result predictions, other ways of incorporating expert opinion is into the model itself [68] or by engaging experts to select model features [62]. For more details on incorporating expert knowledge, please refer to the Appendix.

The most common version of the Brier Score is for a binary prediction, in which case it is equivalent to the Mean Square Error (MSE) and is given by:

$$BS = \frac{1}{N}\sum_{n=1}^{N}(\widehat{y_n} - y_n)^2$$

The BS was deemed inappropriate in the 2017 Soccer Prediction Challenge since it only measures the difference between the predicted and actual scores but does not account for the ordinal nature of the win/draw/loss target variable.

5.3.3 Ranked Probability Score (RPS)

The Ranked Probability Score (RPS) [42] does account for the ordinal nature of the three match outcomes in soccer [29] and was thus deemed more appropriate for model evaluation in the 2017 Soccer Prediction Challenge [39]. The RPS is given by

$$RPS = \frac{1}{r-1}\sum_{i=1}^{r-1}\left(\sum_{j=1}^{i}(p_j - a_j)\right)^2, \tag{5}$$

where r denotes the number of potential match outcomes (e.g., $r = 3$ if there are three possible outcomes: home win, draw, and away win). The RPS value always lies within the interval [0, 1], with an RPS closer to 0 representing a better prediction. The RPS can be averaged across multiple instances as follows:

$$\mathrm{RPS}_{\mathrm{avg}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{RPS}_i. \tag{6}$$

where i denotes the i-th instance in the dataset, RPS_i is the RPS value corresponding to that instance, and N is the total number of instances.

Participants in the 2017 Soccer Prediction Challenge aimed to minimize the average RPS across the $N = 206$ matches in the challenge prediction set (containing matches that participants aimed to predict):

$$\mathrm{RPS}_{\mathrm{avg}} = \frac{1}{206}\sum_{i=1}^{N=206}\mathrm{RPS}_i. \tag{7}$$

5.3.4 Ignorance score (IGN)/Log Loss

The Ignorance score (IGN) was proposed by Good [50], with its foundation lying in information theory. In short, the IGN penalizes predictions with larger logarithmic errors and is given by:

$$IGN = \frac{1}{N}\sum_{n=1}^{N} -(ylog_2(p) + (1-y)log_2(1-p))$$

Where $y \in \{0, 1\}$ and $p = P(y = 1)$. Its value falls in the range $IGN \in [0, \infty)$, with a lower score representing better model performance. If the log base is changed to base e (multiplied by a constant scalar), the IGN is equivalent to log loss.

Based on its properties, IGN has been suggested as being more appropriate than RPS for evaluating the probabilistic forecasts of soccer match result prediction models [119].

5.3.5 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) evaluates numeric prediction models, e.g., that predict the goals scored or goals margin. The RMSE is calculated by first computing the error, the predicted value minus the actual value, and then taking the square root across all instances:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

where N denotes the number of instances, y_i is the actual observed value of the i-th instance, and $\hat{y}_i$ is the predicted value i-th instance.

5.3.6 Scoring rules selection and properties

For numeric prediction problems, RMSE is the most commonly applied scoring rule, while for outcome prediction, the properties of the scoring rule need to be considered. The following three properties are the most commonly described:

- **Proper**: A scoring rule is considered proper when — given inputs $S(P, x)$ with a predicted distribution P, an actual distribution Q, an outcome $x \sim Q$, and scoring rule S — the rule returns a real value $\mathbb{R}$, which we aim to maximize. The expected score is given by $E[S(P, x)] = S(P, Q)$.
  A scoring rule is *strictly proper* if $S(Q, Q) \geq S(P, Q)$ if and only if $Q = P$, which means that the true distribution is given the lowest (most favorable) score and any other distribution receives a higher (less favorable) score.
  On the other hand, a scoring rule is *proper* if $S(Q, Q) \geq S(P, Q)$ for all distributions Q and P. A more theoretical definition can be found in [48].
- **Locality**: A scoring rule is considered local if it only considers the prediction of the observed (target) class. On the other hand, if the scoring rule considers unobserved classes, it is classified as non-local [119].
- **Sensitivity to Distance**: In the context of three-class (home win, draw, away win) prediction, a scoring rule is considered sensitive to distance if it takes into account

the ordinal nature of the outcomes of the match. For example, in soccer, when a team is leading by one goal, it only takes one goal from the opponent to turn the match into a draw, and two goals for the opponent to secure a win. Consequently, the outcome space is ordered based on the goal differences.

Of the above-mentioned scoring rules, the Brier Score, RPS, and IGN/Log Loss are strictly proper, only IGN/Log Loss is local, and only RPS is sensitive to distance.

For the outcome prediction problem, it is essential to use a strictly proper scoring rule to incentivize the model to truthfully report its predictions [48, 16]. The objective is to develop a model that accurately reflects the true distribution of match outcomes.

Regarding the scoring rule for three-class prediction in soccer, Constantinou & Fenton [29] proposed using the RPS because of its sensitivity to distance, i.e., it takes into account the ordering in the outcomes (e.g., a home win is "closer" to a draw than it is to an away win). However, Wheatcroft [119] challenged this perspective, arguing that the sensitivity to distance does not add anything in terms of the usual aims of using scoring rules and that football outcomes may be drawn from an unknown distribution. Therefore, instead of evaluating predictions based on exact outcomes, it was suggested that it is more appropriate to consider the sample distribution of outcomes. In this context, the Ignorance (log loss) Score is a better choice due to its local property, which is more efficient compared to other non-local scoring rules.

While the sensitivity to distance property is still a topic of debate, the scoring rule that is selected should align with the objective of the prediction task.

5.4 Temporal Splitting Methods

The temporal order of matches should be preserved when evaluating results so that upcoming games are predicted based on past games only. One challenge in ML for sports match results prediction is that — unlike many applications of supervised machine learning techniques — the instances representing sports matches are temporally ordered. Furthermore, there is a hierarchy in terms of the way matches are structured, i.e., there are multiple matches in a particular round and multiple rounds that comprise a particular season. As a result, conventional cross-validation, which shuffles instances randomly, cannot be used since future matches may end up being incorrectly used to predict past matches [20, 18, 113]. Performance that appears extremely strong relative to other studies for an equivalent two- or three-class problem (e.g., a study reporting over 80% or 90% accuracy for a three-class problem when most other studies report between 45% to 55%) may be evidence that this mistake has been made in a study.[14] In the machine learning software package WEKA, the "Preserve order for % Split" option in "Classifier evaluation options" can be checked to ensure that the temporal order of instances in a dataset is maintained so that future matches are not incorrectly used to predict historical matches. Code in programming

[14] For example, [102] used 10-fold cross-validation in WEKA and reported obtaining accuracy of 99.56% with a decision tree model.

languages such as Python and R can also be written to ensure that the temporal order of matches is preserved when training and testing models.

Deciding how to split match data into training and validation sets depends on the amount of data that the researcher has on hand, e.g., whether there is one season of match data or multiple seasons. Seasons further in the past become less relevant for predicting matches in the current or future seasons because of, e.g., changes in team rosters and team strengths. However, if player-level features are included, player changes from season to season can be accounted for. The time series nature of soccer match result data should thus be accounted for when researchers evaluate their models, e.g., using cross-validation for time series. Furthermore, any hierarchies that are present in terms of a particular league being made up of seasons that are comprised of rounds, which, in turn, consist of matches, should be considered. If the competition is structured such that teams sometimes play more than one match per round, this should also be taken into account, as should if teams are promoted and relegated at the end of seasons.

> Important

Any match result prediction project must ensure that only matches in the past are being used to predict current or future matches. Therefore, traditional cross-validation, which shuffles data instances randomly, should not be used.

6 Conclusions and Future Directions

Soccer match result prediction is inherently difficult due to soccer being a low-scoring sport with three outcomes and generally highly competitive leagues at the professional level. There is also an element of luck in any sport, including soccer [2], which is, of course, a major part of what attracts fans.

This chapter has aimed to provide a general overview of ML for soccer match result prediction and to act as a resource for researchers conducting future work in this domain. Available datasets, the types of models and features, and approaches to evaluating ML model performance were covered in this chapter, and while the chapter has not aimed to provide an exhaustive review of all studies in the area, it has given a contemporary overview by focusing, to a greater extent, on studies that have been published in recent years. The chapter has also not considered betting strategies to any large degree, which is a significant potential use case for ML match result prediction models in soccer, and researchers are encouraged to also familiarize themselves with the literature in this area if their models will be deployed for this purpose.

As discussed in subsection 2.2, the results from studies that have used the Open International Soccer Database [39] suggest that gradient-boosted tree models, e.g., CatBoost [93] and XGBoost[26], which are applied to feature sets consisting of

soccer-specific ratings (pi-ratings [30] and Berrar ratings [8]), provide state-of-the-art performance for soccer match result prediction in the absence of match features apart from goals scored [8, 59, 96]. It remains unclear, however, whether ensemble methods are the best-performing models on datasets that contain match statistics in addition to goals (e.g., statistics contained in the European Soccer Database, football-data.co.uk) and/or external features, and also whether deep learning models can outperform ensemble models. On other datasets containing match statistics and other types of features, Random Forests have been competitive with [5] and even surpassed the performance gradient-boosted tree models [107, 1, 43]. Thus, comparing the performance of bagging methods (e.g., Random Forest) with boosting methods (e.g., CatBoost) and deep learning models on different datasets with varying features is an interesting avenue for further work. Furthermore, using the recently proposed Generalized Attack Performance (GAP) [118] ratings as model features is yet to be investigated. Rating systems themselves can also be used as predictive models by simply predicting the team with the higher rating as the winner; however, using the ratings as model features has been shown to provide better performance. Despite many sophisticated models having been developed, including hybrid approaches that combine statistical and machine learning models and rating systems [40, 74, 53], predictions derived from bookmaker odds still provide strong baseline performance [100, 5]. There are different ways in which expert knowledge can be incorporated (see the Appendix) into the model itself: through feature selection, or by using the match predictions of experts as a baseline (as mentioned above, former players appear to be better at predicting results [21]). Player or team ratings can be obtained from video games, constructed from performance indicator metrics, or by aggregating player ratings — computed using plus-minus ratings or VAEP — into a team-level rating.

The interpretability of models is of greater importance to certain groups such as sports coaches and performance analysts, and techniques including Random Forest feature importance, SHAP, and models new to the domain such as Alternating Decision Trees (see the Appendix) may be of use in identifying and interpreting performance indicator model features that are most relevant for winning. In the context of deep learning models, which are generally black-box, mimic learning [110] may be worth exploring to aid in their interpretability for match result prediction.

Predicting a different target variable, e.g., expected goals/xG, shot attempts, shots on target, etc., to predict the match result may be superior for prediction than using a three-class win/draw/loss outcome target variable. The best evaluation metric for probabilistic forecasting in this domain remains a subject of debate [30, 119]. Researchers should ensure that their model evaluation adequately accounts for the time series nature of soccer match result data and its hierarchical nature (i.e., match, round, season).

While all existing rating systems account for historical results — as well as accounting for the venue (home/away) and creating separate offensive and defensive ratings in the case of the soccer-specific rating systems — developing new rating systems that incorporate additional factors is an avenue for further research. For instance, rating systems could incorporate information derived from, e.g., event log data, spatiotemporal tracking data, player ratings, player/team attributes and perfor-

mance indicators, passing networks, inter-player chemistry, and even information from social media. Streak and form features could potentially be engineered at both the player (hot shoe) and team levels (form and streaks in terms of match outcomes [5]) and their value as model features compared.

Acknowledgements This chapter was partly supported by JSPS KAKENHI (grant number 20H04075) and JST Presto (grant number JPMJPR20CA).

Appendix

Incorporating Expert Knowledge

The incorporation of domain expert knowledge has been a key area of enquiry for some time. Expert knowledge can be incorporated into the modelling process itself: e.g., Joseph, Fenton & Neil [68] did so in constructing their Bayesian Network. Alternatively, expert knowledge can be incorporated by obtaining expert predictions as a benchmark with which to compare model prediction performance (see also subsection 5.2). Another way in which expert domain knowledge of the sport can be incorporated is by asking experts to select features. Hucaljuk and Rakipović [62] considered 96 matches from the European Champions League and compared the performance of ML models that included features selected by feature selection algorithms with models that included features selected by experts. The feature selection algorithms produced a basic feature set of 20 features in their initial feature set, which they compared with a feature set consisting of this initial set plus the expert-selected features. With their particular models and on their specific dataset, however, the expert-selected features were not found to yield any improvement in model performance. More recently, Beal et al. [6] developed a benchmark dataset and results from Natural Language Processing and ML models for soccer match result prediction, and the performance of these models was compared with statistical models. The authors provided a prediction accuracy baseline that utilizes match statistics and match preview articles written by sports journalists in The Guardian (London, United Kingdom) over six English Premier League seasons.

Elo K-factor Selection Techniques Described Online

The World Football Elo Ratings Website (eloratings.net) [103] provides Elo ratings of national teams, and uses a higher value of K for matches that are of greater importance. For example, $K = 60$ for World Cup finals matches and $K = 20$ for International friendlies. The K-factor is also adjusted based on the goal margin. Specifically, the K-factor is increased by 1/2 for a win by 2 goals, by 3/4 for a win

by 3 goals, and by $\frac{3}{4} + \frac{(M-3)}{8}$ if the team wins by 4 or more goals, where M denotes the goal margin.

The Football Database website (footballdatabase.com) provides Elo ratings for club teams from various leagues around the world, also applying different values of K based on the league or importance of the match.[15] The K-factor is multiplied by a value, G, which accounts for the goal margin. In particular, $G = 1$ if the match was drawn, $G = 1.5$ if the goal margin was 1 or 2 goals, and $G = \frac{11+M}{8}$ if the goal margin was 3 or more, where M again denotes the goals margin.

2023 Soccer Prediction Challenge

A subsequent Soccer Prediction Challenge, using a similar dataset to the 2017 challenge, was held in 2023.[16] However, unlike the 2017 competition, the 2023 version required two tasks. First, predicting match results in terms of the exact goals scored by each team, and second, as per the 2017 challenge, predicting match results by computing the probabilities for a win, draw, and loss. The exact score prediction models were evaluated using the Root-Mean-Squared-Error metric. Similar to the 2017 challenge, studies will be submitted to the Machine Learning (Springer) journal; however, these studies were not available at the time the current chapter was written.

Alternating Decision Trees (ADTrees)

An interesting boosted-tree model that is relatively unexplored in sports result prediction is the Alternating Decision Tree (ADTree) model [46]. As opposed to XGBoost and CatBoost, both of which use gradient boosting and gradient descent optimization, ADTree uses the well-known AdaBoost [47] algorithm. AdaBoost is used to grow the ADTree, with each iteration of AdaBoost adding a branch to the ADTree. The general idea of AdaBoost is to iteratively place more weight on instances that were previously incorrectly classified in a previous iteration. The ADTree algorithm combines the accuracy-increasing benefits of boosting while producing an interpretable decision tree structure as the final ADTree model. Given the interpretability of the ADTree, an enhanced version of ADTree that can compete with the likes of CatBoost would be a great asset to the sports match results prediction domain.[17]

---

[15] See footballdatabase.com/methodology.php for further details.

[16] https://sites.google.com/view/2023soccerpredictionchallenge

[17] Bunker, Yeung, Susnjak, Espie & Fujii [19] recently found that ADTrees performed well in predicting professional ATP tennis match results when applied to the difference in average betting odds across 11 different bookmaker companies (the companies included in the betting odds data on tennis-data.co.uk). Betting odds, as mentioned before, can be considered a rating [122], so this is another example of boosted-tree models, applied to rating features, being effective for sports match result prediction.

References

1. Alfredo, Y.F., Isa, S.M.: Football match prediction with tree based model classification. International Journal of Intelligent Systems and Applications **11**(7), 20–28 (2019)
2. Aoki, R.Y., Assuncao, R.M., Vaz de Melo, P.O.: Luck is hard to beat: The difficulty of sports prediction. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1367–1376 (2017)
3. Arik, S.Ö., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 6679–6687 (2021)
4. Arntzen, H., Hvattum, L.M.: Predicting match outcomes in association football using team ratings and player ratings. Statistical Modelling **21**(5), 449–470 (2021)
5. Baboota, R., Kaur, H.: Predictive analysis and modelling football results using machine learning approach for english premier league. International Journal of Forecasting **35**(2), 741–755 (2019)
6. Beal, R., Middleton, S.E., Norman, T.J., Ramchurn, S.D.: Combining machine learning and human experts to predict match outcomes in football: A baseline model. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 15447–15451 (2021)
7. Berrar, D., Lopes, P., Davis, J., Dubitzky, W.: Guest editorial: special issue on machine learning for soccer. Machine Learning **108**, 1–7 (2019)
8. Berrar, D., Lopes, P., Dubitzky, W.: Incorporating domain knowledge in machine learning for soccer outcome prediction. Machine Learning **108**, 97–126 (2019)
9. Bittner, E., Nußbaumer, A., Janke, W., Weigel, M.: Self-affirmation model for football goal distributions. Europhysics Letters **78**(5), 58002 (2007)
10. Bittner, E., Nußbaumer, A., Janke, W., Weigel, M.: Football fever: goal distributions and non-gaussian statistics. The European Physical Journal B **67**, 459–471 (2009)
11. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)
12. Bransen, L., Van Haaren, J.: Player chemistry: Striving for a perfectly balanced soccer team. arXiv preprint arXiv:2003.01712 (2020)
13. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)
14. Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly Weather Review **78**(1), 1–3 (1950)
15. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN systems **30**(1-7), 107–117 (1998)
16. Bröcker, J., Smith, L.A.: Scoring probabilistic forecasts: The importance of being proper. Weather and Forecasting **22**(2), 382–388 (2007)
17. Brown, A., Reade, J.J.: The wisdom of amateur crowds: Evidence from an online community of sports tipsters. European Journal of Operational Research **272**(3), 1073–1081 (2019)
18. Bunker, R., Susnjak, T.: The application of machine learning techniques for predicting match results in team sport: A review. Journal of Artificial Intelligence Research **73**, 1285–1322 (2022)
19. Bunker, R., Yeung, C., Susnjak, T., Espie, C., Fujii, K.: A comparative evaluation of elo ratings-and machine learning-based methods for tennis match result prediction. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology p. 17543371231212235 (2023)
20. Bunker, R.P., Thabtah, F.: A machine learning framework for sport result prediction. Applied Computing and Informatics **15**(1), 27–33 (2019)
21. Butler, D., Butler, R., Eakins, J.: Expert performance and crowd wisdom: Evidence from english premier league predictions. European Journal of Operational Research **288**(1), 170–182 (2021)
22. Buursma, D.: Predicting sports events from past results towards effective betting on football matches. In: Conference Paper, presented at 14th Twente Student Conference on IT, Twente, Holland, vol. 21. sn (2011)

23. Carloni, L., De Angelis, A., Sansonetti, G., Micarelli, A.: A machine learning approach to football match result prediction. In: HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II 23, pp. 473–480. Springer (2021)
24. Carpita, M., Ciavolino, E., Pasca, P.: Exploring and modelling team performances of the kaggle european soccer database. Statistical Modelling **19**(1), 74–101 (2019)
25. Chen, H.: Neural network algorithm in predicting football match outcome based on player ability index. Advances in Physical Education **9**(4), 215–222 (2019)
26. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
27. Cho, Y., Yoon, J., Lee, S.: Using social network analysis and gradient boosting to develop a soccer win–lose prediction model. Engineering Applications of Artificial Intelligence **72**, 228–240 (2018)
28. Constantinou, A.C.: Dolores: a model that predicts football match outcomes from all over the world. Machine Learning **108**(1), 49–75 (2019)
29. Constantinou, A.C., Fenton, N.E.: Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. Journal of Quantitative Analysis in Sports **8**(1) (2012)
30. Constantinou, A.C., Fenton, N.E.: Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. Journal of Quantitative Analysis in Sports **9**(1), 37–50 (2013)
31. Constantinou, A.C., Fenton, N.E., Neil, M.: Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using bayesian networks. Knowledge-Based Systems **50**, 60–86 (2013)
32. da Costa, I.B., Marinho, L.B., Pires, C.E.S.: Forecasting football results and exploiting betting markets: The case of "both teams to score". International Journal of Forecasting **38**(3), 895–909 (2022)
33. Coulom, R.: Computing "elo ratings" of move patterns in the game of go. ICGA Journal **30**(4), 198–208 (2007)
34. Curley, J.P.: engsoccerdata: English soccer data 1871-2016. R package version 0.1.5 DOI
35. Danisik, N., Lacko, P., Farkas, M.: Football match prediction using players attributes. In: 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), pp. 201–206. IEEE (2018)
36. Decroos, T., Bransen, L., Van Haaren, J., Davis, J.: Actions speak louder than goals: Valuing player actions in soccer. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1851–1861 (2019)
37. Dixon, M.J., Coles, S.G.: Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics) **46**(2), 265–280 (1997)
38. Dmochowski, J.P.: A statistical theory of optimal decision-making in sports betting. PLOS One **18**(6), e0287601 (2023)
39. Dubitzky, W., Lopes, P., Davis, J., Berrar, D.: The open international soccer database for machine learning. Machine Learning **108**, 9–28 (2019)
40. Elmiligi, H., Saad, S.: Predicting the outcome of soccer matches using machine learning and statistical analysis. In: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), pp. 1–8. IEEE (2022)
41. Elo, A.: The rating of chessplayers, past and present (1978)
42. Epstein, E.S.: A scoring system for probability forecasts of ranked categories. Journal of Applied Meteorology (1962-1982) **8**(6), 985–987 (1969)
43. Eryarsoy, E., Delen, D.: Predicting the outcome of a football game: A comparative analysis of single and ensemble analytics methods. In: Proceedings of the 52nd Hawaii International Conference on System Sciences — 2019 (2019)

44. Esme, E., Kiran, M.S.: Prediction of football match outcomes based on bookmaker odds by using k-nearest neighbor algorithm. International Journal of Machine Learning and Computing **8**(1), 26–32 (2018)
45. Forrest, D., Goddard, J., Simmons, R.: Odds-setters as forecasters: The case of english football. International Journal of Forecasting **21**(3), 551–564 (2005)
46. Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: ICML, vol. 99, pp. 124–133 (1999)
47. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55**(1), 119–139 (1997)
48. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association **102**(477), 359–378 (2007)
49. Goes, F., Meerhoff, L., Bueno, M., Rodrigues, D., Moura, F., Brink, M., Elferink-Gemser, M., Knobbe, A., Cunha, S., Torres, R., et al.: Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. European Journal of Sport Science **21**(4), 481–496 (2021)
50. Good, I.J.: Rational Decisions, pp. 365–377. Springer New York, New York, NY (1992). DOI 10.1007/978-1-4612-0919-5_24. URL https://doi.org/10.1007/978-1-4612-0919-5_24
51. Greenhough, J., Birch, P., Chapman, S.C., Rowlands, G.: Football goal distributions and extremal statistics. Physica A: Statistical Mechanics and its Applications **316**(1-4), 615–624 (2002)
52. Groll, A., Ley, C., Schauberger, G., Van Eetvelde, H.: Prediction of the fifa world cup 2018-a random forest approach with an emphasis on estimated team ability parameters. arXiv preprint arXiv:1806.03208 (2018)
53. Groll, A., Ley, C., Schauberger, G., Van Eetvelde, H.: A hybrid random forest to predict soccer matches in international tournaments. Journal of Quantitative Analysis in Sports **15**(4), 271–287 (2019)
54. Guan, S., Wang, X.: Optimization analysis of football match prediction model based on neural network. Neural Computing and Applications pp. 1–17 (2022)
55. Hall, M.A.: Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato (1999)
56. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
57. Holte, R.: Very simple classification rules perform well on most commonly used datasets. Machine Learning **11**, 63–91 (1993)
58. Hubáček, O., Šourek, G., Železnỳ, F.: Exploiting sports-betting market using machine learning. International Journal of Forecasting **35**(2), 783–796 (2019)
59. Hubáček, O., Šourek, G., Železnỳ, F.: Learning to predict soccer results from relational data with gradient boosted trees. Machine Learning **108**, 29–47 (2019)
60. Hubácek, O., Sourek, G., Zelezny, F.: Score-based soccer match outcome modeling–an experimental review. MathSport International (2019)
61. Hubáček, O., Šourek, G., Železnỳ, F.: Forty years of score-based soccer match outcome prediction: an experimental review. IMA Journal of Management Mathematics **33**(1), 1–18 (2022)
62. Hucaljuk, J., Rakipović, A.: Predicting football scores using machine learning techniques. In: 2011 Proceedings of the 34th International Convention MIPRO, pp. 1623–1627. IEEE (2011)
63. Hughes, M.D., Bartlett, R.M.: The use of performance indicators in performance analysis. Journal of Sports Sciences **20**(10), 739–754 (2002)
64. Hvattum, L.M.: A comprehensive review of plus-minus ratings for evaluating individual players in team sports (2019)
65. Hvattum, L.M., Arntzen, H.: Using elo ratings for match result prediction in association football. International Journal of Forecasting **26**(3), 460–470 (2010)
66. Ievoli, R., Palazzo, L., Ragozini, G.: On the use of passing network indicators to predict football outcomes. Knowledge-Based Systems **222**, 106997 (2021)

67. Jain, S., Tiwari, E., Sardar, P.: Soccer result prediction using deep learning and neural networks. In: Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020, pp. 697–707. Springer (2021)
68. Joseph, A., Fenton, N.E., Neil, M.: Predicting football results using bayesian nets and other machine learning techniques. Knowledge-Based Systems **19**(7), 544–553 (2006)
69. Joseph, L.D.: Time series approaches to predict soccer match outcome. Ph.D. thesis, Dublin, National College of Ireland (2022)
70. Kampakis, S., Adamides, A.: Using twitter to predict football outcomes. arXiv preprint arXiv:1411.1243 (2014)
71. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems **30** (2017)
72. Kinalioğlu, İ.H., Kuş, C.: Prediction of football match results by using artificial intelligence-based methods and proposal of hybrid methods. International Journal of Nonlinear Analysis and Applications **14**(1), 2939–2969 (2023)
73. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: D.H. Sleeman, P. Edwards (eds.) Ninth International Workshop on Machine Learning, pp. 249–256. Morgan Kaufmann (1992)
74. Knoll, J., Stübinger, J.: Machine-learning-based statistical arbitrage football betting. KI-Künstliche Intelligenz **34**(1), 69–80 (2020)
75. Kononenko, I.: Estimating attributes: Analysis and extensions of relief. In: F. Bergadano, L.D. Raedt (eds.) European Conference on Machine Learning, pp. 171–182. Springer (1994)
76. Leitner, C., Zeileis, A., Hornik, K.: Is federer stronger in a tournament without nadal? an evaluation of odds and seedings for wimbledon 2009. Austrian Journal of Statistics **38**(4), 277–286 (2009)
77. Leitner, M.C., Daumann, F., Follert, F., Richlan, F.: The cauldron has cooled down: a systematic literature review on home advantage in football during the covid-19 pandemic from a socio-economic and psychological perspective. Management Review Quarterly **73**(2), 605–633 (2023)
78. Ley, C., Wiele, T.V.d., Eetvelde, H.V.: Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. Statistical Modelling **19**(1), 55–73 (2019)
79. Li, J., Zhang, H., Zhao, J., Guo, X., Rihan, W., Deng, G.: Embedded feature selection and machine learning methods for flash flood susceptibility-mapping in the mainstream songhua river basin, china. Remote Sensing **14**(21), 5523 (2022)
80. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems **30** (2017)
81. Maher, M.J.: Modelling association football scores. Statistica Neerlandica **36**(3), 109–118 (1982)
82. Malamatinos, M.C., Vrochidou, E., Papakostas, G.A.: On predicting soccer outcomes in the greek league using machine learning. Computers **11**(9), 133 (2022)
83. Miljković, D., Gajić, L., Kovačević, A., Konjović, Z.: The use of data mining for basketball matches outcomes prediction. In: IEEE 8th International Symposium on Intelligent Systems and Informatics, pp. 309–312. IEEE (2010)
84. Moustakidis, S., Plakias, S., Kokkotis, C., Tsatalas, T., Tsaopoulos, D.: Predicting football team performance with explainable ai: Leveraging shap to identify key team-level performance metrics. Future Internet **15**(5), 174 (2023)
85. Natarajan, S., Khot, T., Kersting, K., Gutmann, B., Shavlik, J.: Gradient-based boosting for statistical relational learning: The relational dependency network case. Machine Learning **86**, 25–56 (2012)
86. Odachowski, K., Grekow, J.: Using bookmaker odds to predict the final result of football matches. In: Knowledge Engineering, Machine Learning and Lattice Computing with Applications: 16th International Conference, KES 2012, San Sebastian, Spain, September 10-12, 2012, Revised Selected Papers 16, pp. 196–205. Springer (2013)

87. Omomule, T., Ibinuolapo, A., Ajayi, O.: Fuzzy-based model for predicting football match results. Int. J. Sci. Res. in Computer Science and Engineering Vol **8**(1) (2020)
88. Ötting, M., Andreas, G.: A regularized hidden markov model for analyzing the 'hot shoe'in football. Statistical Modelling **22**(6), 546–565 (2022)
89. Palinggi, D.A.: Predicting soccer outcome with machine learning based on weather condition. Ph.D. thesis (2019)
90. Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., Giannotti, F.: A public data set of spatio-temporal match events in soccer competitions. Scientific Data **6**(1), 236 (2019)
91. Petretta, M., Schiavon, L., Diquigiovanni, J.: On the dependence in football match outcomes: traditional model assumptions and an alternative proposal. arXiv preprint arXiv:2103.07272 (2021)
92. Pipatchatchawal, C., Phimoltares, S.: Predicting football match result using fusion-based classification models. In: 2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1–6. IEEE (2021)
93. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems **31** (2018)
94. Rahman, M.A.: A deep learning framework for football match prediction. SN Applied Sciences **2**(2), 165 (2020)
95. Randrianasolo, A.S.: Predicting euro games using an ensemble technique involving genetic algorithms and machine learning. In: 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0470–0475. IEEE (2023)
96. Razali, M.N., Mustapha, A., Mostafa, S.A., Gunasekaran, S.S.: Football matches outcomes prediction based on gradient boosting algorithms and football rating system. Human Factors in Software and Systems Engineering **61**, 57 (2022)
97. Razali, N., Mustapha, A., Arbaiy, N., Lin, P.C.: Deep learning for football outcomes prediction based on football rating system. In: AIP Conference Proceedings, vol. 2644. AIP Publishing (2022)
98. Rein, R., Memmert, D.: Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. SpringerPlus **5**(1), 1–13 (2016)
99. Ren, Y., Susnjak, T.: Predicting football match outcomes with explainable machine learning and the kelly index. arXiv preprint arXiv:2211.15734 (2022)
100. Robberechts, P., Davis, J.: Forecasting the fifa world cup–combining result-and goal-based team ability parameters. In: Machine Learning and Data Mining for Sports Analytics: 5th International Workshop, MLSA 2018, Co-located with ECML/PKDD 2018, Dublin, Ireland, September 10, 2018, Proceedings 5, pp. 16–30. Springer (2019)
101. Robberechts, P., Van Haaren, J., Davis, J.: A bayesian approach to in-game win probability in soccer. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 3512–3521 (2021)
102. Rosli, C.M.F.C.M., Saringat, M.Z., Razali, N., Mustapha, A.: A comparative study of data mining techniques on football match prediction. In: Journal of Physics: Conference Series, vol. 1020, p. 012003. IOP Publishing (2018)
103. Runyan, B.: The world football elo rating system. website (1997)
104. Ryall, R., Bedford, A.: An optimized ratings-based model for forecasting australian rules football. International Journal of Forecasting **26**(3), 511–517 (2010)
105. Stefani, R.T.: Football and basketball predictions using least squares. IEEE Transactions on systems, man, and cybernetics **7**(2), 117–21 (1977)
106. Štrumbelj, E., Šikonja, M.R.: Online bookmakers' odds as forecasts: The case of european soccer leagues. International Journal of Forecasting **26**(3), 482–488 (2010)
107. Stübinger, J., Knoll, J.: Beat the bookmaker–winning football bets with machine learning (best application paper). In: Artificial Intelligence XXXV: 38th SGAI International Conference on Artificial Intelligence, AI 2018, Cambridge, UK, December 11–13, 2018, Proceedings 38, pp. 219–233. Springer (2018)

108. Stübinger, J., Mangold, B., Knoll, J.: Machine learning in football betting: Prediction of match results based on player characteristics. Applied Sciences **10**(1), 46 (2019)
109. Sujatha, K., Godhavari, T., Bhavani, N.P.: Football match statistics prediction using artificial neural networks. International Journal of Mathematical and Computational Methods **3** (2018)
110. Sun, X., Davis, J., Schulte, O., Liu, G.: Cracking the black box: Distilling deep sports analytics. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3154–3162 (2020)
111. Surowiecki, J.: The wisdom of crowds. Anchor (2005)
112. Talattinis, K., Kyriakides, G., Kapantai, E., Stephanides, G.: Forecasting soccer outcome using cost-sensitive models oriented to investment opportunities. International Journal of Computer Science in Sport **18**(1) (2019)
113. Tax, N., Joustra, Y.: Predicting the dutch football competition using public data: A machine learning approach. Transactions on Knowledge and Data Engineering **10**(10), 1–13 (2015)
114. Thorp, E.O.: Portfolio choice and the kelly criterion. In: Stochastic Optimization Models in Finance, pp. 599–619. Elsevier (1975)
115. Toda, K., Teranishi, M., Kushiro, K., Fujii, K.: Evaluation of soccer team defense based on prediction models of ball recovery and being attacked: A pilot study. PLOS One **17**(1), e0263051 (2022)
116. Tsokos, A., Narayanan, S., Kosmidis, I., Baio, G., Cucuringu, M., Whitaker, G., Király, F.: Modeling outcomes of soccer matches. Machine Learning **108**, 77–95 (2019)
117. Tümer, A.E., Akyıldız, Z., Güler, A.H., Saka, E.K., Ievoli, R., Palazzo, L., Clemente, F.M.: Prediction of soccer clubs' league rankings by machine learning methods: The case of turkish super league. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology p. 17543371221140492 (2022)
118. Wheatcroft, E.: A profitable model for predicting the over/under market in football. International Journal of Forecasting **36**(3), 916–932 (2020)
119. Wheatcroft, E.: Evaluating probabilistic forecasts of football matches: the case against the ranked probability score. Journal of Quantitative Analysis in Sports **17**(4), 273–287 (2021)
120. Wheatcroft, E.: Forecasting football matches by predicting match statistics. Journal of Sports Analytics **7**(2), 77–97 (2021)
121. Wheatcroft, E., Sienkiewicz, E.: A probabilistic model for predicting shot success in football. arXiv preprint arXiv:2101.02104 (2021)
122. Wunderlich, F., Memmert, D.: The betting odds rating system: Using soccer forecasts to forecast soccer. PLOS One **13**(6), e0198668 (2018)
123. Wunderlich, F., Memmert, D.: A big data analysis of twitter data during premier league matches: do tweets contain information valuable for in-play forecasting of goals in football? Social Network Analysis and Mining **12**, 1–15 (2022)
124. Yeung, C.C., Bunker, R., Fujii, K.: A framework of interpretable match results prediction in football with fifa ratings and team formation. PLOS One **18**(4), e0284318 (2023)
125. Zeileis, A., Leitner, C., Hornik, K.: Probabilistic forecasts for the 2018 fifa world cup based on the bookmaker consensus model. Tech. rep., Working Papers in Economics and Statistics (2018)
126. Zheng, A., Casari, A.: Feature engineering for machine learning: principles and techniques for data scientists. O'Reilly Media, Inc. (2018)
127. Zulkifli, S., Mustapha, A.B., Ismail, S., Razali, N.: Comparative analysis of statistical and machine learning methods for classification of match outcomes in association football. In: Proceedings of the 7th International Conference on the Applications of Science and Mathematics 2021: Sciemathic 2021, pp. 351–365. Springer (2022)

CHAPTER 10: CONCLUDING REMARKS

10.1 Conclusion

This thesis significantly contributes to the emerging field of sports performance analytics through a series of six interconnected publications. This thesis developed frameworks and investigated applications of machine learning and data mining for sports performance analytics. Methods from other domains and contexts were utilised to uncover interpretable key patterns of play and identify performance indicators associated with success at different levels of analysis, and a framework for sports match outcome prediction was developed and applied. These approaches suggested the potential of cross-domain advanced analytics methods and provided evidence of the emergence of a more integrated "sports performance analytics" discipline that combines aspects of sports analytics and sports performance analysis.

Three primary research questions were considered in the course of this work. The first of these questions involved investigating how generalised conceptual frameworks could be established based on critical analysis and synthesis of existing literature. A Sports Results Prediction (SRP-CRISP-DM) framework (Bunker & Thabtah, 2019) was developed as a framework for the application of machine learning for match outcome prediction in sports. This framework arose following a critical analysis of the literature related to the use of artificial neural networks for sports match outcome prediction, which were commonly applied models in early studies in the domain. This work was built upon in Bunker & Susnjak (2022), which investigated a broader range of machine learning techniques and had a more specific focus in terms of the sports investigated (team sports) to provide more in-depth analysis, insights, and conceptual frameworks. Subsequently, the synthesis in Bunker, Yeung, & Fujii (2025) had a more specific focus on machine learning for match outcome prediction in a specific sport, soccer, further refining and developing the enquiry into this domain.

The second question considered how machine learning and data mining methods from other domains could be best leveraged to identify interpretable key patterns. A discriminative sequential pattern mining used for animal trajectory analysis (Sakuma et al., 2019) was shown to generate useful and interpretable event sequence patterns discriminating between labelled sequences at the passage of play level in rugby union (Bunker et al., 2021). In a subsequent study that also investigated rugby union (Bunker & Spencer, 2022), a decision-

rules-based algorithm that had been used in other domains and contexts was shown to be valuable in obtaining interpretable decision rules consisting of performance indicators associated with success at the playoff and group stages of the 2019 RWC tournament. Therefore, both Bunker et al. (2021) and Bunker & Spencer (2022) involved utilising methods that generated interpretable results/patterns.

The final question investigated how machine learning models from other domains and contexts could be employed for interpretable match outcome prediction. The Sports SRP-CRISP-DM framework for sports match result prediction proposed in Bunker & Thabtah (2019) was demonstrated in practice in the context of interpretable machine learning for tennis match result prediction (Bunker et al., 2023). A machine learning model combining the benefits of boosting with an interpretable tree structure, Alternating Decision Trees, was utilised in the context of match outcome prediction in tennis. Its interpretable structure and performance relative to ratings-based methods was favourable (Bunker et al., 2023).

Furthermore, this thesis highlighted the benefits of incorporating advanced analytics techniques, which have traditionally been employed in sports analytics, into sports performance analysis and how the two fields are becoming increasingly interconnected. A framework for sports performance analytics in terms of how it can be defined and a possible workflow was presented in this contextual statement. It was suggested that inter-disciplinary collaboration between researchers in the two disciplines, as well as between researchers in the "parent" disciplines of computer science and statistics and sports science, will continue to drive novelty in the scientific literature, and this will transmit into innovation in practice in this rapidly growing industry.

10.2 Limitations & Future work

There are limitations associated with any research endeavour, and this thesis is no exception. A PhD is designed to develop research skills to be able to carry out independent research effectively, and these skills become honed over time as more knowledge is gained about scientific research. The studies comprising this thesis were carried out over several years, and I believe that my skills as a researcher have indeed improved over this period with each successive publication.

However, with the benefit of hindsight, there are things that I would have done differently had I been starting a similar study from scratch now. For instance, in the survey paper

included in this thesis (Bunker & Susnjak, 2022), if I had known about its existence prior to having written the large majority of the manuscript, it would have been preferable to apply a structured approach to the study in the form of a systematic literature review using an appropriate methodological framework such as the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) framework (Page et al., 2021). To use such a framework would likely have required a narrower focus than team sports because of the number of studies in the area. Therefore, I would carefully consider the scope of a review in a future study to ensure it is sufficiently concise to be able to utilise systematic review frameworks.

As I developed as a researcher, I also became more familiar with the journals available in the field, metrics associated with journal quality, and how to navigate the review process more effectively at higher-quality journals. Therefore, again with the benefit of hindsight, I could perhaps have targeted higher-ranked journals for some of the (particularly early) publications contained in the thesis with the knowledge I have now in terms of the available journals and their quality. This has been valuable experience and gained knowledge for me to aim for such journals in my research going forward.

While a range of sports were covered in the various studies comprising this thesis, a limitation is that generally papers focused on one sport. Of course, some of the techniques employed could also be applied in other sports, given the availability of similar data in that sport. For instance, the MA-Stat-DSM algorithm (Bunker et al., 2024), used to identify important parts of plays from the spatiotemporal data of multiple agents in basketball, could equally be applied to similar data (provided it is available) in other team sports such as soccer, rugby, and ice hockey. Indeed, this is an interesting avenue for further research.

While Elo ratings-based methods were evaluated against machine learning techniques in the context of tennis match outcome prediction in Bunker et al. (2023), it will be interesting in future work to compare their performance against more recently proposed deep learning methods, as well as with the Glicko system (Glickman, 1995) – another extension of Elo ratings, as well as in other sports.

An Australian Academy of Science (2022) report suggested that the professional sports sector currently collects excessive data without fully considering the legal, ethical, and data governance-related ramifications of what can be, in some cases, essentially health

information that is being collected. The report argues for an athlete-centred approach in which athletes are recognised as the owners of their data, as well as the need for specific data collection activities to clearly contribute to a specific organisational objective, for instance, for analysing performance or predicting/preventing injury. The Personal Information Protection and Electronic Documents Act (PIPEDA) and a 1992 decision from the Supreme Court of Canada suggest that athletes retain significant ownership rights over their biometric data (obtained from wearable devices), subject to contracts/collective bargaining agreements (Borden Ladner Gervais LLP, 2022). For the data used in the studies in this thesis, which are mostly derived from in-play events, it is straightforward to justify that this type of data is required for a team's day-to-day objectives of analysing performance. In addition, individual players were not identifiable from the data. In the context of AI systems in sport, Mateus et al. (2024) highlighted the need to comply with data protection laws such as the European General Data Protection Regulations (GDPR), as well as the need for trust to be built in terms of transparent communication with players about data collection practices and their benefits.

One of the studies in this thesis, Bunker et al. (2021) obtained data from a Japan Top League team via the team's performance analyst, from whom written approval was obtained to use the data for research purposes. In reflecting on the content of the Australian Academy of Science report, it would have been perhaps more appropriate to obtain consent from players or, at least, find out whether their playing contracts implied consent. In studies I carry out going forward, I will seek to carefully consider the use of data generated through player actions and/or movements, as well as open datasets that are publicly available, and ensure that the appropriate ethical considerations have been accounted for and necessary steps carried out.

This thesis took a multi-level approach, considering performance at a different level of analysis in each of the studies. However, a limitation of the research is that, although one level of performance was considered in each study, multiple levels were not considered within the same study, which may have allowed for a more holistic analysis. In future work, it could be illuminating to consider multiple levels of analysis within the same study. However, doing so depends on data availability; some datasets are commercially sensitive to professional sports teams and can be challenging to obtain, while others are publicly available. One of the limitations noted in Bunker & Spencer (2022), for example, was that only team-level performance indicator statistics were available – or could be computed from

– the dataset consisting of matches in the 2019 Rugby World Cup. Utilising player-level statistics could have led to more illuminating insights. One source of data that could be made use of in future work is the rugbypy package (https://github.com/seanyboi/rugbypy), which contains statistics at the match, team, and player levels, and according to the package author, was sourced from the ESPN UK website.

Another rich avenue for further research is to develop novel methods that utilise granular event data (e.g., (Bunker et al., 2021) augmented with spatiotemporal tracking data (e.g., Bunker et al., 2024, appendix) by synchronising these two sources of data (e.g., Biermann et al., 2023; Kemp, Wunderlich, & Memmert, 2021). This will become increasingly important as computer vision-based annotation systems become commonplace. In recent years, open source packages such as databallpy (https://databallpy.readthedocs.io/) have been released, which do the heavy lifting in terms of this data augmentation and, although the extent of open data that can be used is often limited by what commercial vendors are willing to provide, the ability for this synchronisation to be performed relatively easily opens the door for research involving the development of novel analytical methods designed to be applied to this type of augmented data. Including features derived from both sources as predictive variables in match outcome prediction models while minimising information loss associated with aggregation is also an exciting avenue for further work.

There is also the opportunity to develop and apply methods that already exist in major sports, such as soccer and basketball, in more minor sports, such as rugby union. An expected points model akin to expected goals in soccer could be devised for rugby union (a recent preprint by Fitzpatrick & Nolan (2024) considered expected points in rugby but does not take a machine learning approach). Such an expected points model may be helpful for decision support, for instance, when a team is awarded a penalty in the opposition half and needs to decide an appropriate course of action.

In subsection 2.3.3, the long-term performance of teams over multiple seasons was discussed briefly. Some studies have attempted to identify ‘eras’ when a sport is relatively stable and when eras change, for example, through an evolution in match play or rule changes (Woods, Robertson, & Collier, 2017; Young, Luo, Gastin, Tran, & Dwyer, 2019). This has also been investigated in specific facets of sports, such as home advantage in soccer (Jacklin, 2005). It could be interesting to extend this type of analysis to identifying eras for specific teams in a sport.

Finally, deep learning is a recent and promising field for further investigation in sports performance analytics, however, it wasn't applied in any of the studies contained in this thesis. However, given the desirability of interpretability in sports and the black-box nature of deep learning models, these will need to be used with explainable (xAI) techniques to determine the effects of particular model features on performance-related outcomes.

REFERENCES

Agrawal R., & Srikant R. (1995). Mining Sequential Patterns. In: *Proceedings of the Eleventh International Conference on Data Engineering*. (pp. 3–14). IEEE.

Alamar, B.C. & Mehrotra, V. (2011). *"Beyond 'Moneyball': The Rapidly Evolving World of Sports Analytics, Part I*", Analytics Magazine (September 2011).

Alamar, B.C. (2013). *Sports Analytics: A Guide for Caches, Managers and Other Decision Makers*. Columbia University Press (New York).

Angelini, G., Candila, V., & De Angelis, L. (2022). Weighted Elo rating for tennis match predictions. *European Journal of Operational Research*, 297(1), 120-132.

Aoki, R. Y., Assuncao, R. M., & Vaz de Melo, P. O. (2017). Luck is hard to beat: The difficulty of sports prediction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1367–1376.

Apostolou, K., & Tjortjis, C. (2019, July). Sports Analytics algorithms for performance prediction. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1-4). IEEE.

Australian Academy of Science (2022). Getting Ahead of the Game: Athlete Data in Professional Sport.

Behravan, I., & Razavi, S. M. (2021). A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing*, 25(3), 2499-2511.

Bekkers, J., & Dabadghao, S. (2019). Flow motifs in soccer: What can passing behavior tell us?. *Journal of Sports Analytics*, 5(4), 299-311.

Bennett, M., Bezodis, N. E., Shearer, D. A., & Kilduff, L. P. (2021). Predicting performance at the group-phase and knockout-phase of the 2015 Rugby World Cup. *European Journal of Sport Science*, 21(3), 312-320.

Berri, D., Butler, D., Rossi, G., Simmons, R., & Tordoff, C. (2024). Salary determination in professional football: empirical evidence from goalkeepers. *European Sport Management Quarterly*, 24(3), 624-640.

Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, 108, 97-126.

Berrar, D., Lopes, P., & Dubitzky, W. (2024). A data-and knowledge-driven framework for developing machine learning models to predict soccer match outcomes. *Machine Learning*, 1-40.

Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S., & Matthews, I. (2014, December). Identifying team style in soccer using formations learned from spatiotemporal

tracking data. In *2014 IEEE International Conference on Data Mining Workshop* (pp. 9-14). IEEE.

Biermann, H., Komitova, R., Raabe, D., Müller-Budack, E., Ewerth, R., & Memmert, D. (2023). Synchronization of passes in event and spatio-temporal soccer data. *Scientific Reports*, 13(1), 15878.

Bilek, G., & Ulas, E. (2019). Predicting match outcome according to the quality of opponent in the English premier league using situational variables and team performance indicators. *International Journal of Performance Analysis in Sport,* 19(6), 930-941.

Borden Ladner Gervais LLP. (2022, September 29). *Ownership of athlete biometric data in Canadian sports.* https://www.blg.com/en/insights/2022/09/ownership-of-athlete-biometric-data-in-canadian-sports

Borms, J. (2008). *Directory of Sport Science: A Journey Through Time: The Changing Face of ICSSPE*. Champaign, IL: Human Kinetics.

Bunker, R. (2022). The Bogey Phenomenon in Sport. *IX Mathsport International 2022 Proceedings*.

Bunker, R. P., & Spencer, K. (2022). Performance indicators contributing to success at the group and play-off stages of the 2019 Rugby World Cup. *Journal of Human Sport and Exercise, 17*(3), 683-698.

Bunker, R. P., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research, 73*, 1285-1322.

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics, 15*(1), 27-33.

Bunker, R. P., Fujii, K., Hanada, H., & Takeuchi, I. (2021). Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: an application to rugby union. *PLOS One, 16*(9), e0256329.

Bunker, R., Fujii, K., Hanada, H., & Takeuchi, I. (2021). Supervised sequential pattern mining for identifying important patterns of play in rugby. In *Proceedings of the 8th MathSport International Conference*, June 2021. Online.

Bunker, R., Yeung, C., Fujii, K. (2025). Machine Learning for Soccer Match Result Prediction. In: Blondin, M.J., Fister Jr., I., Pardalos, P.M. (eds) *Artificial Intelligence, Optimization, and Data Sciences in Sports*. Springer Optimization and Its Applications, vol 218. Springer, Cham. https://doi.org/10.1007/978-3-031-76047-1_2

Bunker, R. P., Yeung, C., & Fujii, K. (2024). An expected wins approach using Fisher's Exact Test to identify the bogey effect in sports: An application to tennis. *Journal of Sport & Exercise Science*, 8(1), 43-54.

Bunker, R. P., Yeung, C., Susnjak, T., Espie, C., & Fujii, K. (2023). A comparative evaluation of Elo ratings-and machine learning-based methods for tennis match result prediction. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, 17543371231212235.

Bunker, R., Yeung, C., Susnjak, T., Espie, C., & Fujii, K. (2023). A comparison of the performance of Elo ratings and machine learning techniques for match result prediction in tennis. Paper presented at the *14th European Sport Economics Association (ESEA) Conference* 2023, Cork, Ireland.

Butler, R. J., Smith, M., & Irwin, I. (1993). The performance profile in practice. *Journal of Applied Sport Psychology*, 5(1), 48-63.

Butterworth, A., O'Donoghue, P., & Cropley, B. (2013). Performance profiling in sports coaching: a review. *International Journal of Performance Analysis in Sport, 13*(3), 572-593.

Camerino, O. F., Chaverri, J., Anguera, M. T., & Jonsson, G. K. (2012). Dynamics of the game in soccer: Detection of T-patterns. *European Journal of Sport Science*, 12(3), 216-224.

Cavus, M., & Biecek, P. (2022, October). Explainable expected goal models for performance analysis in football analytics. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 1-9). IEEE.

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Choi, H., O'Donoghue, P., & Hughes, M. (2007). An investigation of inter-operator reliability tests for real-time analysis system. *International Journal of Performance Analysis in Sport*, 7(1), 49-61.

Chmait, N., & Westerbeek, H. (2021). Artificial intelligence and machine learning in sport research: An introduction for non-data scientists. Frontiers in sports and active living, 3, 682287.

Cioppa, A., Giancola, S., Deliege, A., Kang, L., Zhou, X., Cheng, Z., ... & Van Droogenbroeck, M. (2022). Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3491-3502).

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge, New York. https://doi.org/10.4324/9780203771587

Cohen, W. W. (1995). Fast effective rule induction. In *Machine learning proceedings 1995* (pp. 115-123). Morgan Kaufmann.

Cohen, W. W., & Singer, Y. (1996, August). Learning to query the web. In *AAAI Workshop on Internet-Based Information Systems* (pp. 16-25).

Cole, J., & Martin, A. J. (2018). Developing a winning sport team culture: organizational culture in theory and practice. *Sport in Society*, 21(8), 1204-1222.

Constantinou, A. C., & Fenton, N. E. (2012). Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, 8(1).

Constantinou, A. C., & Fenton, N. E. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, 9(1), 37-50.

Constantinou, A. C. (2019). Dolores: a model that predicts football match outcomes from all over the world. *Machine Learning*, 108(1), 49-75.

Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., & Wang, L. (2023). SportsMOT: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9921-9931).

Cust, E. E., Sweeting, A. J., Ball, K., & Robertson, S. (2019). Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance. *Journal of Sports Sciences*, 37(5), 568-600.

Dale, G. A., & Wrisberg, C. A. (1996). The use of a performance profiling technique in a team setting: Getting the athletes and coach on the “same page”. *The Sport Psychologist*, 10(3), 261-277.

Davis, J., Bransen, L., Devos, L., Jaspers, A., Meert, W., Robberechts, P., ... & Van Roy, M. (2024). Methodology and evaluation in sports analytics: Challenges, approaches, and lessons learned. *Machine Learning*, 113(9), 6977-7010.

Decroos, T., Van Haaren, J., & Davis, J. (2018, July). Automatic discovery of tactics in spatio-temporal soccer match data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 223-232).

Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019, July). Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1851-1861).

Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting*, 28(2), 543-552.

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2), 265-280.

Dolejš, J., & Jureček, M. (2022). Interpretability of machine learning-based results of malware detection using a set of rules. In *Artificial Intelligence for Cybersecurity* (pp. 107-136). Cham: Springer International Publishing.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78-87.

Dubitzky, W., Lopes, P., Davis, J., & Berrar, D. (2019). The open international soccer database for machine learning. *Machine Learning*, 108, 9-28.

Elstak, I., Salmon, P., & McLean, S. (2024). Artificial intelligence applications in the football codes: A systematic review. *Journal of Sports Sciences*, 42(13), 1184-1199.

Epstein, E.S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology* (1962-1982), 8 (6), 985–987.

Eryarsoy, E., & Delen, D. (2019). Predicting the outcome of a football game: A comparative analysis of single and ensemble analytics methods. *Proceedings of the 52nd Hawaii International Conference on System Sciences 2019.*

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-37.

Ferrero, C. A., Alvares, L. O., Zalewski, W., & Bogorny, V. (2018, April). Movelets: Exploring relevant subtrajectories for robust trajectory classification. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (pp. 849-856).

Fitzpatrick, B., & Nolan, D. The application of expected points in rugby union: proposal of a novel framework. sportRxiv preprint. https://doi.org/10.51224/SRXIV.444

Fortune Business Insights (2024, September 16). *Sports Analytics Market Size, Share & Industry Analysis, By Deployment (Cloud and On-premise), By Type (On-field and Off-field), By Solution (Video Analytics, Bio Analytics, Smart Wearable Technology, and Others), By Technology (AI, Big Data, and Others), By End-user (Team and Individual), and Regional Forecast, 2024-2032*. Fortune Business Insights. https://www.fortunebusinessinsights.com/sports-analytics-market-102217

Fournier-Viger, P., Gomariz, A., Campos, M., & Thomas, R. (2014). Fast vertical mining of sequential patterns using co-occurrence information. In *Advances in Knowledge*

Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part I 18 (pp. 40-52). Springer International Publishing.

Fournier-Viger, P., Lin, J. C. W., Kiran, R. U., Koh, Y. S., & Thomas, R. (2017). A survey of sequential pattern mining. *Data Science and Pattern Recognition*, 1(1), 54-77.

Franks, I. M., & Miller, G. (1991). Training coaches to observe and remember. *Journal of Sports Sciences*, 9(3), 285-297.

Frencken, W., Lemmink, K., Delleman, N., & Visscher, C. (2011). Oscillations of centroid position and surface area of soccer teams in small-sided games. *European Journal of Sport Science*, 11(4), 215-223.

Freund, Y., & Mason, L. (1999, June). The alternating decision tree learning algorithm. In *ICML* (Vol. 99, pp. 124-133).

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119-139.

Fujii, K., Yamada, K., Kono, R., Zhang, Z., & Bunker, R (2025). Machine learning-based analysis of multi-agent trajectories in basketball. To appear as a book chapter in *Artificial Intelligence and Machine learning in Sports Science* (Springer).

García, J., Ibáñez, S. J., De Santos, R. M., Leite, N., & Sampaio, J. (2013). Identifying basketball performance indicators in regular season and playoff games. *Journal of Human Kinetics*, 36(1), 161-168.

Geurkink, Y., Boone, J., Verstockt, S., & Bourgois, J. G. (2021). Machine learning-based identification of the strongest predictive variables of winning and losing in Belgian professional soccer. *Applied Sciences*, *11*(5), 2378.

Giles, B., Peeling, P., Kovalchik, S., & Reid, M. (2023). Differentiating movement styles in professional tennis: A machine learning and hierarchical clustering approach. *European Journal of Sport Science*, 23(1), 44-53.

Glickman, M. E. (1995). The glicko system. *Boston University*, *16*(8), 9.

Goel, A., & Mallick, B. (2015). Customer purchasing behavior using sequential pattern mining technique. International Journal of Computer Applications, 119(1).

Goes, F. R., Brink, M. S., Elferink-Gemser, M. T., Kempe, M., & Lemmink, K. A. (2021). The tactics of successful attacks in professional association football: large-scale spatiotemporal analysis of dynamic subgroups using position tracking data. *Journal of Sports Sciences*, 39(5), 523-532.

Goes, F., Kempe, M., & Lemmink, K. (2019). *Predicting match outcome in professional Dutch football using tactical performance metrics computed from position tracking data* (No. 993). EasyChair.

Goes, F. R., Kempe, M., Meerhoff, L. A., & Lemmink, K. A. (2019). Not every pass can be an assist: a data-driven model to measure pass effectiveness in professional soccer matches. *Big Data*, 7(1), 57-70.

Good, I. J. 1992. *"Rational Decisions." In Breakthroughs in Statistics*, 365–77. New York: Springer.10.1007/978-1-4612-0919-5_24

GQ. (2024, May 27). *The Greatest Seasons in Team Sports History*. GQ. https://www.gq.com.au/gq-sports/greatest-seasons-team-sports-history/image-gallery/66ccac8f1e94f935bf015894bcf7cf6f

Gomez-Ruano, M. A., Ibáñez, S. J., & Leicht, A. S. (2020). Performance analysis in sport. *Frontiers in Psychology*, 11, 611634.

Gonzales, A., Guruswamy, G., & Smith, S. R. (2023). Synthetic data in health care: A narrative review. PLOS Digital Health, 2(1), e0000082.

Green, S. (2012). *Assessing the Performance of Premier League Goalscorers.* Statsbomb. https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/

Groll, A., Ley, C., Schauberger, G., & Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of quantitative analysis in sports*, 15(4), 271-287.

Grosskreutz, H., & Rüping, S. (2009). On subgroup discovery in numerical domains. *Data Mining and Knowledge Discovery*, 19, 210-226.

Grunz, A., Memmert, D., & Perl, J. (2012). Tactical pattern recognition in soccer games by means of special self-organizing maps. *Human Movement Science*, 31(2), 334-343.

Gudmundsson, J., & Horton, M. (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys* (CSUR), 50(2), 1-34.

Gyarmati, L., Kwak, H., & Rodriguez, P. (2014). Searching for a unique style in soccer. arXiv preprint arXiv:1409.0308.

Hubáček, O., Šourek, G., & Železný, F. (2019a). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, 35(2), 783-796.

Hubáček, O., Šourek, G., & Železnỳ, F. (2019b). Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, 108, 29–47.

Hughes, M. D., & Bartlett, R. M. (2002). The use of performance indicators in performance analysis. *Journal of Sports Sciences, 20*(10), 739-754.

Hughes, M., & Bartlett, R. (2019). What is performance analysis?. In *Essentials of Performance Analysis in Sport, 3rd Ed.* (pp. 11-19). Routledge.

Hughes, A., Barnes, A., Churchill, S. M., & Stone, J. A. (2017). Performance indicators that discriminate winning and losing in elite men's and women's Rugby Union. *International Journal of Performance Analysis in Sport*, 17(4), 534-544.

Hughes, M., Evans, S., & Wells, J. (2001). Establishing normative profiles in performance analysis. *International Journal of Performance Analysis in Sport, 1*(1), 1-26.

Humphreys, B. R. (2002). Alternative measures of competitive balance in sports leagues. Journal of Sorts Economics, 3(2), 133-148.

Hrovat, G., Fister Jr, I., Yermak, K., Stiglic, G., & Fister, I. (2015). Interestingness measure for mining sequential patterns in sports. *Journal of Intelligent & Fuzzy Systems*, 29(5), 1981-1994.

Hsu, Y. C. (2021). Using convolutional neural network and candlestick representation to predict sports match outcomes. *Applied Sciences*, 11(14), 6594.

Hvattum, L. M. (2020). Offensive and defensive plus–minus player ratings for soccer. *Applied Sciences*, 10(20), 7345.

IBM (2024). What is Machine learning (ML)? https://www.ibm.com/topics/machine-learning

Jabbar, M. A., Deekshatulu, B. L., & Chndra, P. (2014, November). Alternating decision trees for early diagnosis of heart disease. In *International Conference on Circuits, Communication, Control and Computing* (pp. 322-328). IEEE.

Jacklin, P. B. (2005). Temporal changes in home advantage in English football since the Second World War: What explains improved away performance?. *Journal of Sports Sciences*, 23(7), 669-679.

Jayal, A., McRobert, A., Oatley, G., & O'Donoghue, P. (2018). *Sports Analytics: Analysis, Visualisation and Decision Making in Sports Performance*. Routledge (London, UK).

Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7), 544-553.

Kahn, J. (2003). Neural network prediction of NFL football games. World Wide Web electronic publication, 9–15.

Karim, M. R., Shajalal, M., Graß, A., Döhmen, T., Chala, S. A., Boden, A., ... & Decker, S. (2023, October). Interpreting black-box machine learning models for high

dimensional datasets. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics* (DSAA) (pp. 1-10). IEEE.

Kharrat, T., McHale, I. G., & Peña, J. L. (2020). Plus–minus player ratings for soccer. *European Journal of Operational Research*, 283(2), 726-736.

Klemp, M., Wunderlich, F., & Memmert, D. (2021). In-play forecasting in football using event and positional data. *Scientific Reports*, 11(1), 24139.

Kovalchik, S. A. (2023). Player tracking data in sports. *Annual Review of Statistics and Its Application*, 10(1), 677-697.

La Puma, I., & de Castro Giorno, F. A. (2017, December). Ontology-based data mining approach for judo technical tactical analysis. In *The Third International Conference on Computing Technology and Information Management* (ICCTIM2017) (p. 90).

Lago-Peñas, C., Lago-Ballesteros, J., & Rey, E. (2011). Differences in performance indicators between winning and losing teams in the UEFA Champions League. *Journal of Human Kinetics*, 27(2011), 135-146.

Lames, M., & McGarry, T. (2007). On the search for reliable performance indicators in game sports. *International Journal of Performance Analysis in Sport*, 7(1), 62-79.

Le Duy, V. N., Sakuma, T., Ishiyama, T., Toda, H., Arai, K., Karasuyama, M., ... & Takeuchi, I. (2020). Stat-DSM: Statistically discriminative sub-trajectory mining with multiple testing correction. *IEEE Transactions on Knowledge and Data Engineering*, 34(3), 1477-1488.

Leitner, C., Zeileis, A., & Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, 26(3), 471-481.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton & Company.

Li, Y., Ma, R., Gonçalves, B., Gong, B., Cui, Y., & Shen, Y. (2020). Data-driven team ranking and match performance analysis in Chinese Football Super League. *Chaos, Solitons & Fractals*, 141, 110330.

Link, D. (2018). *Data analytics in professional soccer*. Springer Vieweg, Wiesbaden, 10, 978-3.

Link, D., & Lames, M. (2009). Sport informatics: Historical roots, interdisciplinarity and future developments. *International Journal of Computer Science in Sport*, 8(2), 68-87.

Link, D., & Lames, M. (2014). An introduction to sport informatics. In *Computer science in sport* (pp. 1-17). Routledge (London, UK).

Liu, H., Hopkins, W., Gómez, A. M., & Molinuevo, S. J. (2013). Inter-operator reliability of live football match statistics from OPTA Sportsdata. *International Journal of Performance Analysis in Sport*, 13(3), 803-821.

Liu, K. Y., Lin, J., Zhou, X., & Wong, S. T. (2005, December). Boosting alternating decision trees modeling of disease trait information. In *BMC Genetics* (Vol. 6, pp. 1-6). BioMed Central.

Lord F., Pyne D.B., Welvaert M., Mara J.K. (2020) Methods of performance analysis in team invasion sports: A systematic review. *Journal of Sports Sciences,* 38: 2338-2349.

Lundberg, S. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

Mackay, L., Jones, B., van Rensburg, D. C. C. J., Hall, F., Alexander, L., Atkinson, K., ... & Whitehead, S. (2023). Consensus on a netball video analysis framework of descriptors and definitions by the netball video analysis consensus group. *British Journal of Sports Medicine*, 57(8), 441-449.

Mackenzie, R., & Cushion, C. (2013). Performance analysis in football: A critical review and implications for future research. *Journal of Sports Sciences*, 31(6), 639-676.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3), 109-118.

Martin, L. (2015). Is socioeconomic status a contributing factor to tennis players' success? *Journal of Medicine and Science in Tennis*, 20(3), 116-121.

Martin, L. (2016). *Sports performance measurement and analytics: The science of assessing performance, predicting future outcomes, interpreting statistical models, and evaluating the market value of athletes*. FT Press.

Martin, L. (2019). Sports science data protocol. *Sport Exerc Med Open J*, 5, 36-41.

Mateus, N., Abade, E., Coutinho, D., Gómez, M. Á., Peñas, C. L., & Sampaio, J. (2024). Empowering the Sports Scientist with Artificial Intelligence in Training, Performance, and Health Management. Sensors, 25(1), 139.

Me, E., & Unold, O. (2011). Machine learning approach to model sport training. *Computers in Human Behavior*, 27(5), 1499-1506.

Miller, T. W. (2015). *Sports Analytics and Data Science: Winning the Game with Methods and Models*. FT press.

Memmert, D., Lemmink, K. A., & Sampaio, J. (2017). Current approaches to tactical performance analyses in soccer using position data. *Sports Medicine*, 47(1), 1-10.

Mohaghegh, S. D. (2021, April 1). Traditional statistics vs. artificial intelligence and machine learning. *Journal of Petroleum Technology*. https://jpt.spe.org/traditional-statistics-vs-artificial-intelligence-and-machine-learning

Morgulev, E., & Lebed, F. (2024). Beyond key performance indicators: Theoretical-methodological discussion of performance analysis (sports analytics) research. *German Journal of Exercise and Sport Research*, 1-6.

Moustakidis, S., Plakias, S., Kokkotis, C., Tsatalas, T., & Tsaopoulos, D. (2023). Predicting football team performance with explainable AI: Leveraging SHAP to identify key team-level performance metrics. *Future Internet*, 15(5), 174.

Murphy, A. (1969). On the "ranked probability score". *Journal of Applied Meteorology* , 8, 988-989.

Murphy, A. (1970). The ranked probability score and the probability score: A comparison. *Monthly Weather Review*, 98, 917-924.

Nakagawa, K., Suzumura, S., Karasuyama, M., Tsuda, K., & Takeuchi, I. (2016, August). Safe pattern pruning: An efficient approach for predictive pattern mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1785-1794).

O'Donoghue, P. (2005). Normative profiles of sports performance. *International Journal of Performance Analysis in Sport*, 5(1), 104-119.

O'Donoghue, P. (2009). *Research Methods for Sports Performance Analysis.* Routledge (Abingdon, UK).

O'Donoghue, P. (2013). Sports performance profiling. In *Routledge Handbook of Sports Performance Analysis* (pp. 127-139). Routledge (London, UK).

O'Donoghue, P. (2014). *An Introduction to Performance Analysis of Sport*. Routledge (London, UK).

Ortega, B. P., & Olmedo, J. M. J. (2017). Application of motion capture technology for sport performance analysis. *Retos: nuevas tendencias en educación física, deporte y recreación*, (32), 241-247.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Palut, Y., & Zanone, P. G. (2005). A dynamical analysis of tennis: Concepts and data. *Journal of Sports Sciences*, 23(10), 1021-1032.

Pantzalis, V. C., & Tjortjis, C. (2020, July). Sports analytics for football league table and player performance prediction. In *2020 11th International Conference on Information, Intelligence, Systems and Applications* (IISA (pp. 1-8). IEEE.

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019). PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5), 1-27.

Parim, C., Güneş, M. Ş., Büyüklü, A. H., & Yıldız, D. (2021). Prediction of match outcomes with multivariate statistical methods for the group stage in the UEFA Champions League. *Journal of Human Kinetics*, 79(1), 197-209.

Parmar, N., James, N., Hughes, M., Jones, H., & Hearne, G. (2017). Team performance indicators that predict match outcome and points difference in professional rugby league. *International Journal of Performance Analysis in Sport*, 17(6), 1044-1056.

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., ... & Hsu, M. C. (2004). Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11), 1424-1440.

Pham, B. T., Tien Bui, D., & Prakash, I. (2017). Landslide susceptibility assessment using bagging ensemble based alternating decision trees, logistic regression and J48 decision trees methods: a comparative study. *Geotechnical and Geological Engineering*, 35, 2597-2611.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.

Purucker, M. C. (1996). Neural network quarterbacking. *IEEE Potentials*, 15(3), 9-15.

Rahman, M. A. (2020). A deep learning framework for football match prediction. *SN Applied Sciences*, 2(2), 165.

Razali, M. N., Mustapha, A., Mostafa, S. A., & Gunasekaran, S. S. (2022). Football matches outcomes prediction based on gradient boosting algorithms and football rating system. *Human Factors in Software and Systems Engineering*, 61, 57.

Read, B., & Edwards, P. (1992). *Teaching Children to Play Games*. Leeds: White Line Publishing.

Reed, D., & O'Donoghue, P. (2005). Development and application of computer-based prediction methods. *International Journal of Performance Analysis in Sport*, 5(3), 12-28.

Reep, C., & Benjamin, B. (1968). Skill and chance in association football. *Journal of the Royal Statistical Society*. Series A (General), 131(4), 581-585.

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, 5, 1-13.

Ren, Y., & Susnjak, T. (2022). Predicting football match outcomes with explainable machine learning and the kelly index. arXiv preprint arXiv:2211.15734.

Robertson, S., Back, N., & Bartlett, J. D. (2016). Explaining match outcome in elite Australian Rules football using team performance indicators. *Journal of Sports Sciences*, 34(7), 637-644.

Robertson, S., Gupta, R., & McIntosh, S. (2016). A method to assess the influence of individual player performance distribution on match outcome in team sports. *Journal of Sports Sciences*, 34(19), 1893–1900. https://doi.org/10.1080/02640414.2016.1142106

Rommers, N., Rössler, R., Verhagen, E., Vandecasteele, F., Verstockt, S., Vaeyens, R., ... & Witvrouw, E. (2020). A machine learning approach to assess injury risk in elite youth football players. *Medicine and Science in Sports and Exercise*, 52(8), 1745-1751.

Roulston, M. S., & Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6), 1653-1660.

Ryall, R., & Bedford, A. (2010). An optimized ratings-based model for forecasting Australian Rules football. *International Journal of Forecasting*, 26(3), 511-517.

Sakuma, T., Nishi, K., Kishimoto, K., Nakagawa, K., Karasuyama, M., Umezu, Y., ... & Takeuchi, I. (2019). Efficient learning algorithm for sparse subsequence pattern-based classification and applications to comparative animal trajectory data analysis. *Advanced Robotics*, 33(3-4), 134-152.

Salvemini E., Fumarola F., Malerba D., & Han J. (2011). Fast sequence mining based on sparse id-lists. In *International Symposium on Methodologies for Intelligent Systems* (pp. 316–325). Springer, Berlin, Heidelberg.

Sampaio, J., Lago-Peñas, C., & Gómez, M. A. (2013). Brief exploration of short and mid-term timeout effects on basketball scoring according to situational variables. European Journal of Sport Science, 13(1), 25-30.

Sampaio, J., & Leite, N. (2013). Performance indicators in game sports. In *Routledge handbook of sports performance analysis* (pp. 115-126). Routledge.

Sarmento, H., Marcelino, R., Anguera, M. T., CampaniÇo, J., Matos, N., & LeitÃo, J. C. (2014). Match analysis in football: a systematic review. *Journal of Sports Science*s, 32(20), 1831-1843.

Schwartz, B., & Barsky, S. F. (1977). The home advantage. *Social Forces*, 55(3), 641-661.

Scott, A., Uchida, I., Ding, N., Umemoto, R., Bunker, R., Kobayashi, R., ... & Fujii, K. (2024). TeamTrack: A Dataset for Multi-Sport Multi-Object Tracking in Full-pitch Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3357-3366).

Severini, T. A. (2014). Analytic methods in sports: Using mathematics and statistics to understand data from baseball, football, basketball, and other sports (1st Edition). *Chapman and Hall/CRC*.

Shearer, C. (2000) The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5, 13-22.

Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research*, 110, 102817.

Simpson, I., Beal, R.J., Locke, D., & Norman, T.J. (2022). Seq2event: Learning the language of soccer using transformer-based match event prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 3898–3908).

s.v. "performance, n., sense 1.b". (2023, September). Retrieved from Oxford English Dictionary, .: https://doi.org/10.1093/OED/2675064267

Szymanski, S. (2020). Sport analytics: Science or alchemy?. *Kinesiology Review*, 9(1), 57-63.

Srikant R. & Agrawal R. (1996). Mining sequential patterns: Generalizations and performance improvements. In: *International Conference on Extending Database Technology* (pp. 1–17). Springer, Berlin, Heidelberg.

Stats Perform a (n.d.). Opta Data. Stats Perform. https://www.statsperform.com/opta/

Stats Perform b. (n.d.). Opta event definitions. Stats Perform. https://www.statsperform.com/opta-event-definitions/

Stübinger, J., Mangold, B., & Knoll, J. (2019). Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, 10(1), 46.

Tax, N., & Joustra, Y. (2015). Predicting the Dutch football competition using public data: A machine learning approach. *Transactions on knowledge and data engineering*, 10(10), 1-13.

Thabtah, F., & Kamalov, F. (2017). Phishing detection: a case analysis on classifiers with rules using machine learning. *Journal of Information & Knowledge Management*, 16(04), 1750034.

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA game result prediction using feature analysis and machine learning. *Annals of Data Science*, 6(1), 103-116.

Thomas, G., Gade, R., Moeslund, T. B., Carr, P., & Hilton, A. (2017). Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding*, 159, 3-18.

Wang, K., Li, H., Gong, L., Tao, J., Wu, R., Fan, C., ... & Cui, P. (2020, October). Match tracing: A unified framework for real-time win prediction and quantifiable performance evaluation. In *Proceedings of the 29th ACM International conference on information & knowledge management* (pp. 2781-2788).

Wang, K., Xu, Y., & Yu, J. X. (2004, November). Scalable sequential pattern mining for biological sequences. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management* (pp. 178-187).

Watson, N., Hendricks, S., Stewart, T., & Durbach, I. (2021). Integrating machine learning and decision support in tactical decision-making in rugby union. *Journal of the Operational Research Society*, 72(10), 2274-2285.

West, L. J., & Hankin, R. K. (2008). Exact tests for two-way contingency tables with structural zeros. *Journal of Statistical Software*, 28, 1-19.

Webb, P. I., Pearson, P. J., & Forrest, G. (2006). *Teaching Games for Understanding (TGfU) in Primary and Secondary Physical Education*.

Wheatcroft, E. (2020). A profitable model for predicting the over/under market in football. *International Journal of Forecasting*, 36(3), 916-932.

Wheatcroft, E. (2021). Evaluating probabilistic forecasts of football matches: the case against the ranked probability score. *Journal of Quantitative Analysis in Sports*, 17(4), 273-287.

White, R., Palczewska, A., Weaving, D., Collins, N., & Jones, B. (2021). Sequential movement pattern-mining (SMP) in field-based team-sport: A framework for quantifying spatiotemporal data and improve training specificity? *Journal of Sports Sciences*, 40(2), 164–174. https://doi.org/10.1080/02640414.2021.1982484

Wilkens, S. (2021). Sports prediction and betting models in the machine learning age: The case of tennis. *Journal of Sports Analytics*, 7(2), 99-117.

Witten, I. H., Frank, E., Hall, M. A., (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann, CA, USA.

Woods, C. T., Robertson, S., & Collier, N. F. (2017). Evolution of game-play in the Australian Football League from 2001 to 2015. *Journal of Sports Sciences*, 35(19), 1879-1887.

Wu, Y., Ke, Y., Chen, Z., Liang, S., Zhao, H., & Hong, H. (2020). Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. *Catena*, 187, 104396.

Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of Experimental Orthopaedics*, 8, 1-15.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Vogelbein, M., Nopp, S. and Hökelmann, A. (2014). Defensive transition in soccer–are prompt possession regains a measure of success? A quantitative analysis of German Fußball-Bundesliga 2010/2011. *Journal of Sports Sciences*, 32(11), 1076-1083.

Ye, L., & Keogh, E. (2011). Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data Mining and Knowledge Discovery*, 22, 149-182.

Yeung, C., & Bunker, R. (2023). An events and 360 data-driven approach for extracting team tactics and evaluating performance in football. *Statsbomb Conference 2023 Research Paper*.

Yeung, C., Bunker, R., & Fujii, K. (2023). A framework of interpretable match results prediction in football with FIFA ratings and team formation. *Plos One*, 18(4), e0284318.

Yeung, C., Bunker, R., & Fujii, K. (2024). Unveiling Multi-Agent Strategies: A Data-Driven Approach for Extracting and Evaluating Team Tactics from Football Event and Freeze-Frame Data. *Journal of Robotics and Mechatronics*, 36(3), 603-617.

Yeung, C., Bunker, R., Umemoto, R., & Fujii, K. (2024). Evaluating soccer match prediction models: a deep learning approach and feature optimization for gradient-boosted trees. *Machine Learning*, 1-24.

Yeung, C. C., Sit, T., & Fujii, K. (2023). Transformer-based neural marked spatio temporal point process model for football match events analysis. arXiv preprint arXiv:2302.09276.

Young, C., Luo, W., Gastin, P., Tran, J., & Dwyer, D. (2019). Modelling match outcome in Australian Football: improved accuracy with large databases. *International Journal of Computer Science in Sport*, 18(1), 80-92.

Yücebaş, S. C. (2022). A deep learning analysis for the effect of individual player performances on match results. *Neural Computing and Applications*, 34(15), 12967-12984.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42, 31-60.

Zhang, Z., Bunker, R., Takeda, K., & Fujii, K. (2023). Multi-agent deep-learning based comparative analysis in basketball. In *Proceedings of the 37th National Conference of the Japanese Society for Artificial Intelligence (2023)*, pp. 3U1IS304-3U1IS304. The Japanese Society for Artificial Intelligence, 2023.

Zhang, Q., Zhang, X., Hu, H., Li, C., Lin, Y., & Ma, R. (2022). Sports match prediction model for training and exercise using attention-based LSTM network. *Digital Communications and Networks*, 8(4), 508-515.

Zheng, Y. (2015). Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology* (TIST), 6(3), 1-41.

Zhu, P., & Sun, F. (2020). Sports athletes' performance prediction model based on machine learning algorithm. In *International Conference on Applications and Techniques in Cyber Intelligence ATCI 2019: Applications and Techniques in Cyber Intelligence 7* (pp. 498-505). Springer International Publishing.

Zimbalist, A. S. (2002). Competitive balance in sports leagues: An introduction. Journal of Sports Economics, 3(2), 111-121.

Ziyi, Z., Bunker, R., Takeda, K., & Fujii, K. (2023). Multi-agent deep-learning based comparative analysis of team sport trajectories. *IEEE Access*, 11, 43305-43315.

APPENDIX

A1. Classifying types of sports

In the main text, invasion sports and striking/fielding sports are referred to. A taxonomy of sports is shown in Figure 12.

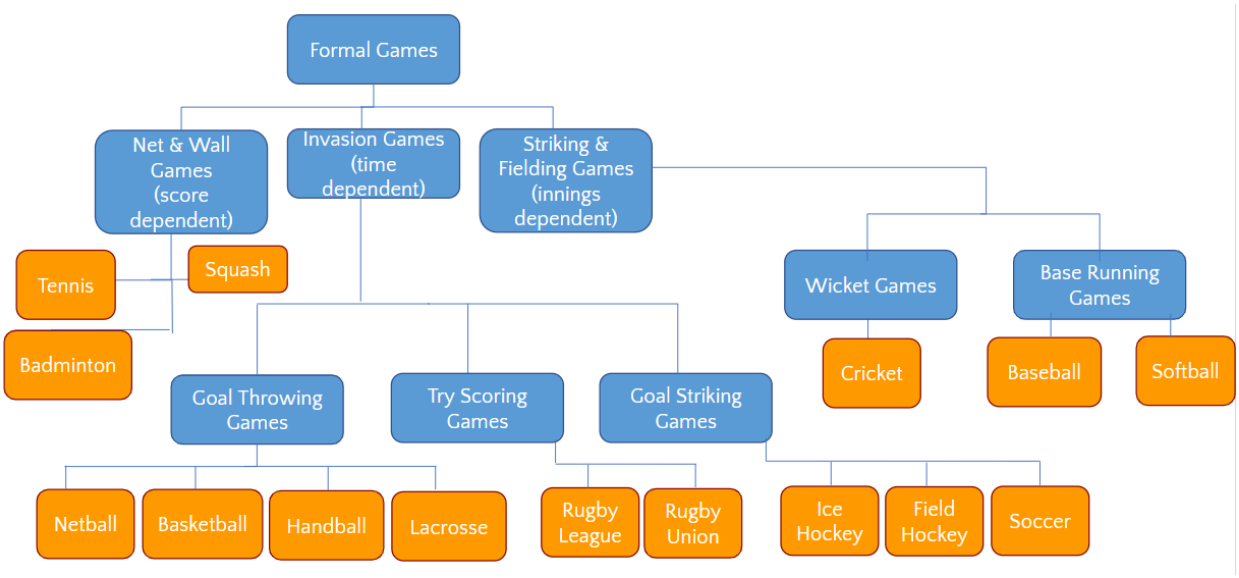


Figure 12: Taxonomy of formal games. *Based on the categorisations provided by Read and Edwards (1992) and Hughes and Bartlett (2002)*

A2. Additional related publications

The following two papers were accepted for publication after commencing the enrolment period of the PhD by prior publication. These two studies are connected to the six included papers and are, therefore, included in this appendix. The two papers are shown in green in Figure 13 in relation to the main text publications based on the level of performance investigated.

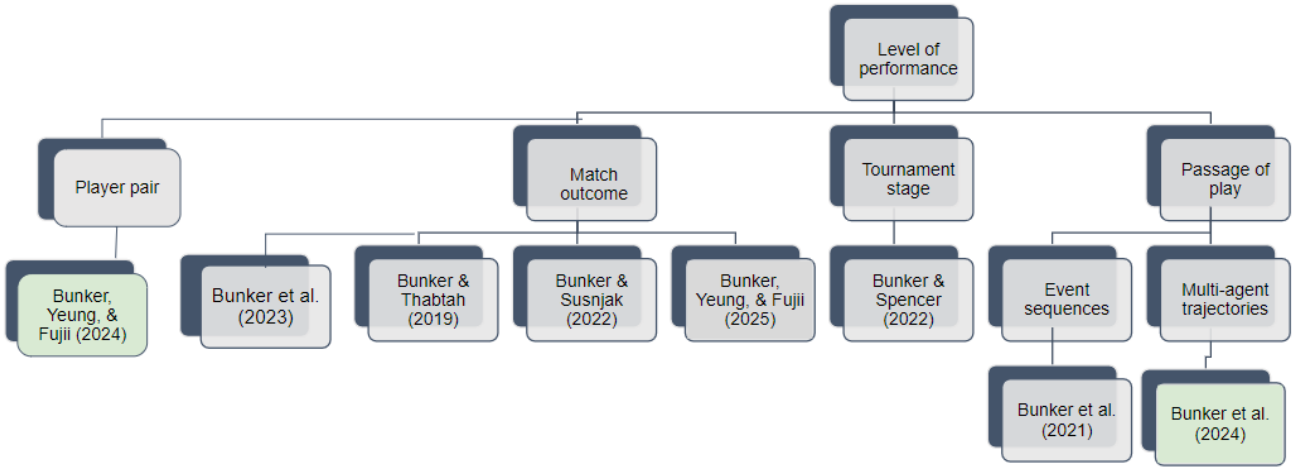


Figure 13: Additional related publications. *Positioning of the two related publications in relation to the main-text publications in terms of level of performance investigated*

Table 2: Additional publications connected to thesis. *The DOIs/URLs, the level at which performance was analysed, study type, and methodologies employed of two additional studies connected to those contained in this thesis but not included in the main text.*

Study Title	**DOI/URL**	**Study Type**	**Level at which performance was analysed**	**Methodologies considered**
Bunker, R; Le Duy, V; Tabei, Y; Takeuchi, I; Fujii, K (2024); Multi-agent statistically discriminative sub-trajectory mining and an application to NBA basketball. *Journal of Quantitative Analysis in Sports.* 2024 Sep 23.	https://doi.org/10.1515/jqas-2023-0039	Development of a novel method and application	Passage of play outcome (labelled with effective/ineffective)	Discriminative sub-trajectory mining
Bunker, R. P., Yeung, C., & Fujii, K. (2024). An expected wins approach using Fisher's Exact Test to identify the bogey effect in sports: An application to tennis. *Journal of Science in Sport and Exercise*, 8(1), 43-54.	https://doi.org/10.36905/jses.2024.01.06	Development of new method & application	Player pair outcomes (which reflects all historical match outcomes between pairs of players, i.e., the win count for each player in a player pair)	Fisher's exact test, Elo ratings

A2.1 PUBLICATION 7: "AN EXPECTED WINS APPROACH USING FISHER'S EXACT TEST TO IDENTIFY THE BOGEY EFFECT IN SPORTS: AN APPLICATION TO TENNIS"

Bunker, R. P., Yeung, C., & Fujii, K. (2024). An expected wins approach using Fisher's Exact Test to identify the bogey effect in sports: An application to tennis. *Journal of Sport & Exercise Science*, 8(1), 43-54.

I obtained a summer research scholarship during the summer of 2015- 2016. I worked with Dr Robin Hankin, a senior lecturer in statistics at Auckland University of Technology at the time, on a sports statistics project about the bogey team phenomenon in sports. A statistical method called the Aylmer test (West & Hankin, 2008), an extension of Fisher's Exact test that accounts for structural zeros, was utilised to identify bogey teams using National Rugby League (NRL) match result data. I then presented a conference paper with another approach that used runs tests at the MathSport International 2022 conference (Bunker, 2022).

Building on these prior studies, the current study considered performance at the player pair level, which incorporates the historical wins for each player in a player pair. The study proposed a statistical procedure to identify bogey players in tennis who tend to beat a particular opposition with regularity despite seemingly being of similar ability. A publicly available dataset containing ATP (men's) and WTA (women's) data were used and also partitioned into datasets consisting of Grand Slam tournaments and non-Grand Slam matches for these two leagues.

The proposed method iterates over all unique player pairs, applying Fisher's Exact test to a contingency table containing expected win and actual win distributions for historical matches between the players in the player pair. To compute expected wins, the betting odds- or Elo ratings-implied probabilities for each player are aggregated over all matches between the players. Suppose Fisher's Exact test result is statistically significant (actual wins and expected wins do not follow the same distribution), and the expected wins and actual win counts are contradictory. In that case, the bogey effect exists between the two players.

The results found that the bogey player effect exists in professional men's and women's tennis, but it is rare and even rarer in Grand Slams. Expected wins obtained using betting

odds-implied win probabilities were found to be more closely aligned with actual wins than Elo ratings, resulting in fewer bogey player pairs being identified with betting odds.

Bunker, R. P., Yeung, C., & Fujii, K. (2024). An expected wins approach using Fisher's Exact Test to identify the bogey effect in sports: An application to tennis. Journal of Sport & Exercise Science, 8(1), 43-54. https://doi.org/10.36905/jses.2024.01.06
Shared with permission from the the Journal of Sport & Exercise Science / Sport & Exercise New Zealand (SESNZ).

An expected wins approach using Fisher's Exact Test to identify the bogey effect in sports: An application to tennis

Rory P. Bunker[1*], Calvin Yeung[1], Keisuke Fujii[1,2,3]

[1]*Graduate School of Informatics, Nagoya University, Japan*
[2]*RIKEN Advanced Intelligence Project, Japan*
[3]*PRESTO Japan Science and Technology Agency, Japan*

Keywords:
Tennis
Bogey
Betting odds
Fisher's Exact Test
Elo ratings

A B S T R A C T

In sports, so-called "bogey" players or teams tend to beat a particular opposition with regularity despite being of similar ability. Although the existence of bogey players is widely discussed and debated among sports fans and the media, methods that could be used to identify the bogey effect have received little attention in the literature. This study proposes a statistical procedure to identify bogey players using a publicly available men's Association of Tennis Professionals (ATP) and Women's Tennis Association (WTA) dataset, which is also split into Grand Slam and non-Grand Slam matches. The proposed method iterates over all unique player pairs and applies Fisher's Exact Test to a contingency table containing expected wins and actual win distributions for historical matches between the players in the player pair. To compute expected wins, betting odds- or Elo ratings-implied probabilities for each player are aggregated over all matches between the player pair for each player. If the Fisher's Exact Test result is statistically significant (that is, actual wins and expected wins do not follow the same distribution), and the expected wins and actual win counts are contradictory, we suggest that the bogey effect exists between the two players. The obtained results suggest that the bogey player effect exists in professional tennis but is rare (and even rarer in Grand Slams), and expected wins obtained using betting odds-implied win probabilities more closely matched actual wins than Elo ratings, which resulted in fewer bogey player pairs being identified with betting odds. The number of bogey player pairs identified is intuitively found to be inversely related to the predictability of matches.

1. Introduction

The existence of the bogey effect, in which players (or teams in the case of team sports) tend to beat a particular opposition with regularity despite being of similar ability, is widely discussed among sports fans and the media. For instance, France at one point was considered the bogey team of New Zealand in Rugby World Cup tournaments (Bruce, 2014), while in soccer, Italy was considered the bogey team of Germany (Wilms, 2013). While tennis, which is considered in the current study, appears to have had less attention in the media than soccer, there has been some discussion of potential bogey players in online forums and in articles (Niall, 2013; Wood, 2017). Loosely speaking, a bogey player/team, or "Angstgegener" (translated as "feared opponent") as it is known in the German language, tends to habitually beat another specified player/team despite appearing to be of equal, or even lesser, strength on paper. Although the bogey phenomenon has been mentioned in a small number of academic studies in, for example, education (Bruce, 2014) and sociology (Chiweshe, 2021; Poulton, 2004), and in doctoral theses (Awerbuch, 2009; Wilms, 2014), it has been largely unexplored in the sports science and sports statistics disciplines.

Related to the bogey effect are the concepts of streaks, form, hot hand, stability (non-stationarity), autocorrelation, and "hot and cold nights". Streaks can be considered both in terms of

*Corresponding Author: Rory Bunker, Graduate School of Informatics, Nagoya University, Japan, rorybunker@gmail.com

individual player actions (e.g., home runs in baseball, three-pointers in basketball) or match-winning streaks. Form, also known as the "hot hand" phenomenon in basketball, assumes that future outcomes can be determined, at least partly, based on the most recent outcomes and that players or teams having successful streaks impact their future successes (Ayton & Fischer, 2004; Bar-Eli, Avugos, & Raab, 2006). Carlson and Shu (2007) found that across five diverse studies, including one related to shooting in basketball, the third repeated event within a sequence is critical to the subjective belief that a streak is happening. The most common techniques used in such analyses have been Wald-Wolfowitz Runs Tests and autocorrelation tests (Carlson & Shu, 2007; Peel & Clauset, 2015; Raab, 2012; Stone, 2012). Hales (1999) argued that autocorrelation, the correlation between the outcomes of consecutive events, and non-stationarity, the probability of success fluctuating over time, should be considered separately because the two concepts represent different underlying mechanisms of the hot hand effect. Specifically, (positive) autocorrelation, which is often measured using the correlation coefficient or the runs test, suggests that success in one event increases the likelihood of success in a subsequent event, thus indicating a hot hand effect. If autocorrelation is present, it may indicate that performance is influenced by recent success/failure, and a measure of positive association between shot outcomes may be appropriate. Non-stationarity may be caused by several factors (e.g., form, fatigue, or external factors such as opponent performance) and the chi-square test can be used to detect changes in the probability of success over time. Non-stationarity indicates the existence of fluctuations in player performance, and a time-varying ability parameter (e.g., time-varying Bradley-Terry parameters as per Cattelan, Varin, & Firth, 2013) may be appropriate to model the data. Steeger, Dulin, and Gonzalez (2021) distinguished between streaks and momentum; the former referring to observed sequences of events each of which may or may not have dependence between them, while momentum suggests that a dependence exists between events that are similar. In their seminal paper, Gilovich, Vallone, and Tversky (1985) refer to basketball shooters having "hot" and "cold" nights (i.e., strong and poor performance, respectively), and analysed whether stability exists across matches in terms of shooters having more hot or cold nights than would be expected by chance; that is, how the variability in match shooting percentages that is observed compares with the expected variability according to a player's record overall. The "gambler's fallacy", also known more generally as negative recency, is the belief that in sequences comprised of binary random events, runs of a specific outcome will be balanced by corrective action; that is, a tendency for the other outcome (Estes, 1964 and Ayton & Fisher, 2004, as cited in Steeger, Dulin, & Gonzalez, 2021). Positive recency, also known as the hot-hand fallacy, is the inclination to predict future outcomes the same as recent outcomes (Ayton & Fischer, 2004, as cited in Steeger, Dulin, & Gonzalez, 2021). Baboota and Kaur (2019) engineered streak and time-weighted streak features for soccer match result prediction that consider the results of a single, and a form feature that considers the results between specific pairs of teams. At first glance, it can be tempting to attribute a long streak of wins to a particular team to the bogey effect. Tottenham, for example, won only one game out of 37 against Chelsea between 1990 and 2006, and Watford did not beat Manchester City in the 30-year period from 1989 to 2019. However, it may have been the case that these results may all have been expected; thus, the question then becomes how expectedness can be accounted for. The media tend to use the "bogey" term relatively loosely, without providing additional contextual information that would help determine whether such results are actually unexpected. This study attempts to clarify this by introducing a bogey player identification method that uses Fisher's Exact Test to determine whether the match results between a particular pair of tennis players deviate from what would be expected, given the betting odds or Elo ratings of the two players in each player pair.

Prior studies that have focused on bogey effect identification specifically have attempted to use statistical techniques including the Wald-Wolfowitz runs test (Bunker, 2022) and the Aylmer test (Hankin & Bunker, 2016), using data from Tennis and Rugby League. Using runs tests, however, also tended to identify result sequences evenly split between unexpected wins and unexpected losses, which is not indicative of the bogey effect, and the Aylmer test, since it was applied to the match results of all pairs simultaneously, was subject to the multiple comparisons issue. In the current study, betting odds, which were used to identify unexpected results in Bunker (2022), as well as Elo ratings are used to compute expected wins for each player pair. In particular, (implied) win probabilities for both betting odds and Elo ratings are aggregated to calculate expected wins for each player in the player pair. Leitner, Zeileis, and Hornik (2009) proposed a bookmaker consensus model that aggregates bookmaker expectations into a prediction for tennis match results. In the proposed method, the odds that we use in the dataset already represent the average across multiple bookmakers, and we aggregate the odds-implied win probabilities of each player in the player pair. Fisher's Exact Test (FET) is then applied to a contingency table consisting of the computed expected wins distribution and the actual win distribution for the player pair. The bogey effect is considered to exist between two players if FET yields a statistically significant result, and the expected wins and actual wins contradict. The method proposed circumvents the multiple comparisons problem since it is applied iteratively to each unique player pair, and match results and betting odds are independent since odds are determined by bookmakers prior to matches and thus represent their assessment of the probabilities of the match outcomes, and bets placed on certain outcomes do not affect match results, while match result is determined purely by the teams' or players' in-match performance. While odds can provide insights into bookmaker expectations, in the absence of match-fixing, they do not have any effect on match results. As well as proposing a novel method for bogey effect identification, the method is demonstrated on publicly available datasets consisting of 33,976 men's Association of Tennis Professionals (ATP) matches from 2005 to 2020 and 27,094 Women's Tennis Association (WTA) matches from 2007 to 2020, as well as subsets of the original datasets comprising only Grand Slam and non-Grand Slam matches. The analysis presented in this paper is also relevant in the context of sports psychology in that some players (teams) struggling more against a particular opposition player (team) may be a psychological phenomenon.

We hypothesise that the betting odds are more accurate for Grand Slam matches since, first, bookmakers would carry out more research and perform more modelling before setting the odds for the match outcome, and second, the betting volume is likely to be higher on Grand Slam matches compared to non-

Grand Slam matches, and this additional information from betting volume increases the accuracy of the odds. Since betting odds accuracy is inversely related to the number of bogey players identified by odds in our proposed method, both of these factors would contribute to fewer bogey players being identified.

On the other hand, Elo ratings are affected only by match results and are updated dynamically over time based on these match outcomes. Form and strength would also be accounted for by bookmakers, and they may even use Elo ratings or other ratings in their models to set odds, however, this is only a subset of the information they would use to set betting odds.

Through a regression-based analysis of 51,881 tennis matches, Barrutiabengoa, Corredor, and Muga (2022) found that, even after controlling for surprise factor uncertainty and the amount of media attention, the prices that bookmakers quote are higher for women's matches than men's, which suggests that the betting volume (and, therefore, betting odds accuracy) on women's matches could be lower compared to men's. Vaughan Williams, Liu, Dixon, and Gerrard (2021) compared the performance of betting odds, rankings, standard and surface-specific Elo ratings, and weighted rating composites, including and excluding the betting odds, in predicting men's and women's professional tennis matches and found that betting odds performed well in general, and standard Elo ratings performed well for women's tennis. The authors found that Elo and betting odds performed better than rankings, which supports the use of these two variables in our proposed method.

Kovalchik (2016) found that the accuracy of predictive models when predicting match results is markedly different for lower-ranked and top-ranked players, finding that match outcome prediction models are 10% to 20% less accurate for matches among lower-ranked players than matches among top-ranked players. Yue, Chou, Hsieh, and Hsiao (2022) showed that win probability with respect to ranking difference fluctuates to a greater degree (i.e., predictability decreases) when the ranking difference increases (Figure 2, Yue et al., 2022) due to smaller number of samples at larger ranking differences (Figure 3, Yue et al., 2022). The findings of Yue et al. (2022) imply then that matches between top-ranked players have a small ranking difference (Elo rating difference) and are thus easier to predict, while matches between top-ranked and lower-ranked players, which have a large ranking (Elo rating) difference, are more difficult to predict. In this study, we partition the original dataset into Grand Slam and non-Grand Slam matches. Grand Slams generally consist of matches among top-ranked players, thus, based on the findings of Kovalchik (2016) and Yue et al. (2022), we would expect our proposed method to identify fewer bogey player pairs for the Grand Slam match dataset compared to the non-Grand Slam match dataset.

Elo ratings, which are updated over time based on historical match results, do not incorporate factors such as the court surface and what hand the players play with, while betting odds do incorporate such factors through the odds set by bookmakers supplemented by fan knowledge reflected in betting volume, which bookmakers use to tweak their initial odds. We would expect that using Elo ratings in the proposed method will generally identify more bogey players than betting odds.

2. Methods

2.1. Data

2.1.1. Datasets

Publicly available data from professional ATP (men's) and WTA (women's) tennis was sourced for this study. In particular, we use the same datasets used by Angelini, Candila, and De Angelis (2022), which were originally sourced from the website tennis-data.co.uk. Among many other variables, the dataset contains match data from ATP and WTA tournaments and Grand Slams, as well as bookmaker odds, player rankings, and ATP/WTA player points. The datasets contain 33,976 men's ATP matches from 2005 to 2020 and 27,094 women's WTA matches from 2007 to 2018. The ATP and WTA datasets were downloaded as RData files from "Appendix C. Supplementary materials" in the paper by Angelini, Candila, & De Angelis (2022). An R script was then created to clean the data using the `clean()` function in the welo R package (Candila, 2023) and to convert and export the cleaned data to a CSV file. The datasets were passed through the welo R package's `clean()` function, which reduced the number of matches in the final dataset. This data-cleaning function reduced the original number of matches from 38,868 to 33,976 for the ATP dataset and from 30,706 to 27,094 for the WTA dataset.[1]

For further analysis, the ATP and WTA datasets consisting of all matches were each divided into two additional datasets consisting of Grand Slam matches and non-Grand Slam matches (Figure 1), and the proposed method will also be applied to these four additional datasets.

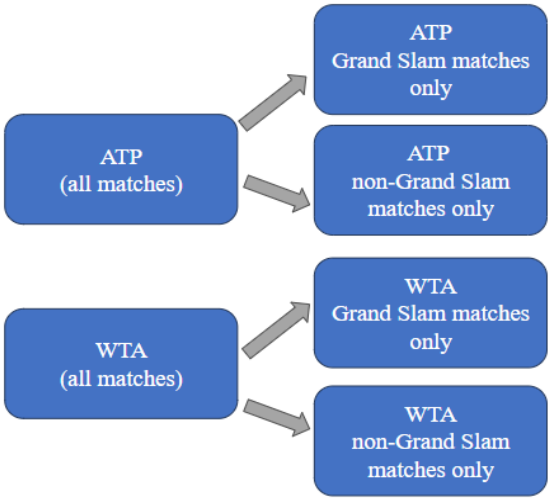


Figure 1: Datasets used upon which the proposed method is applied.

2.1.2. Descriptive statistics

[1] The `clean()` function performs ten cleaning operations, which are described under the clean function details on page 4 in the welo package manual: https://cran.r-project.org/package=welo. The default options of the function were used.

To compare the temporal variability of betting odds (winner, loser, and combined) and Elo ratings, we plot the coefficient of variation (CV), which is calculated by dividing the standard deviation by the mean and is indicative of the level of dispersion around the mean and accounts for the different scale of variance/standard deviation of odds and Elo ratings, of each for both the ATP and WTA datasets (Figure 2A and Figure 3A). As can be seen, the variability of Elo ratings is much lower overall compared to betting odds. It is notable from Figure 3 that the variability of betting odds for women's tennis declined over the time period 2005 and 2020. The variability of Elo ratings declined for both ATP and WTA in the latter period, however, WTA Elo rating variability declined from 2013 onwards, whereas ATP Elo rating variability declined later – from 2016 onwards. The mean Elo rating generally increased over the time period for both ATP and WTA. Another noticeable feature of Figures 2C and 3C is that the mean betting odds and betting odds CV for the WTA, especially for the loser odds, declined over the sample period. On the other hand, the mean betting odds and betting odds CV for the ATP had no discernible trend over the sample period. This perhaps indicates that bookmakers, on account of an evening in the level of competition in the WTA, as evidenced by the decline in Elo rating CV from 2013 onwards, began to lower the price on the likely losers in WTA matches. The remaining figures for the Grand Slam and non-Grand Slam datasets for the ATP and WTA are in the Appendix (Supplementary Figures 1 to 4).

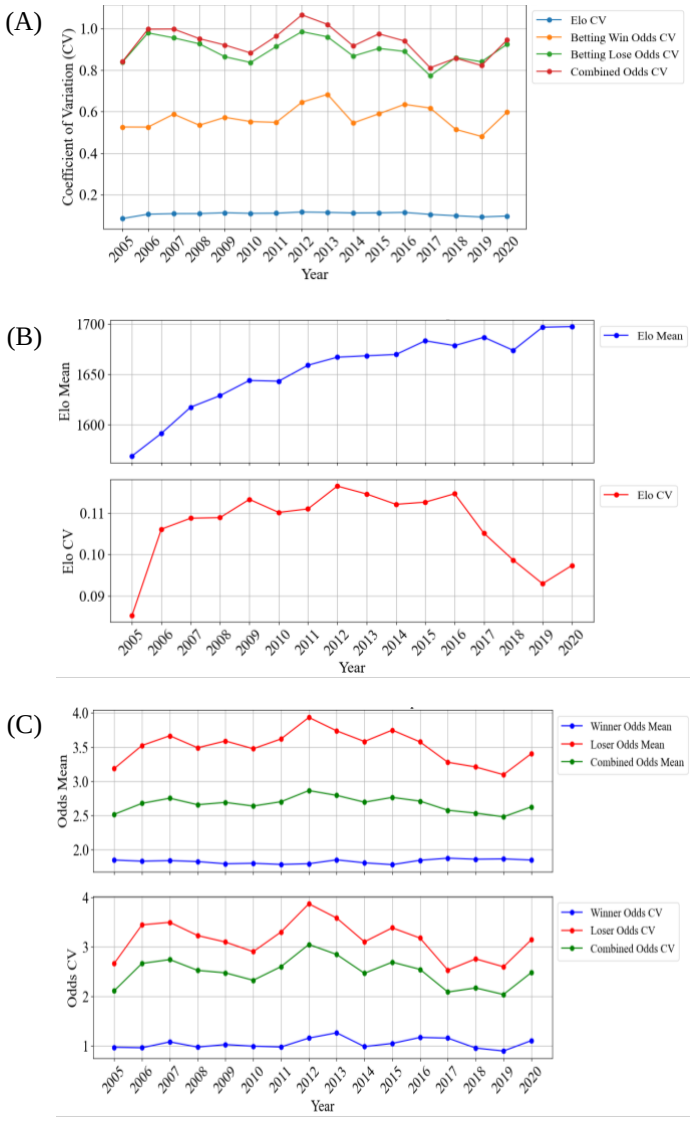


Figure 2: Coefficient of Variation (CV) of Betting Odds and Elo Ratings for each year (A), as well as the more granular view of the mean and CV for each year of Elo rating and Betting Odds (B and C, respectively) for the ATP dataset.

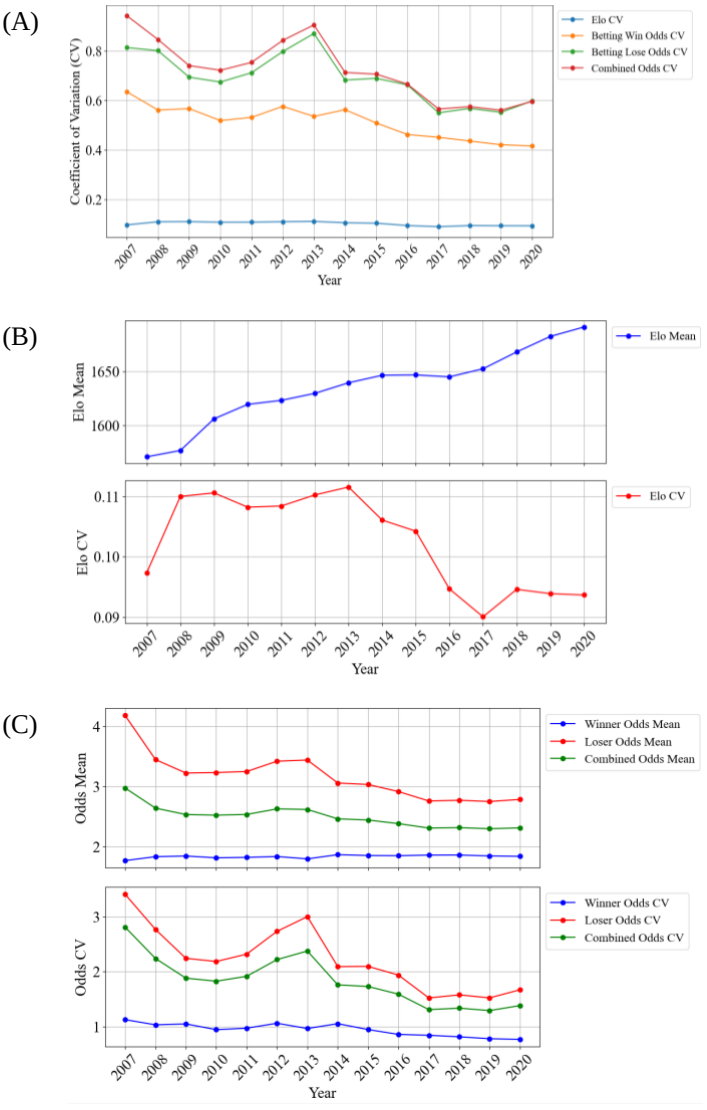


Figure 3. Coefficient of Variation (CV) of Betting Odds and Elo Ratings for each year (A), as well as the more granular view of the mean and CV for each year of Elo rating and Betting Odds (B and C, respectively) for the WTA dataset.

2.2. Proposed method

The proposed method consists of three steps, which are depicted in Figure 4 and are outlined in the following three subsections.

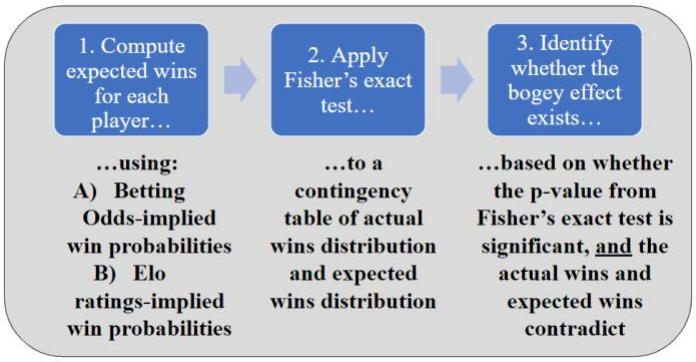


Figure 4: Three steps of the proposed method.

2.2.1. Computing expected wins

As mentioned, expected wins are determined using two approaches: using betting odds- and Elo ratings-implied probabilities. So that the win probabilities for each player in each match add to one, the betting odds-implied win probabilities are obtained by dividing the reciprocal of the decimal betting odds by a normalization factor (also known as the over-round), which is simply the sum of the two decimal betting odd reciprocals. Mathematically, the win probabilities for each player, i and j, in match t, can thus be derived from the decimal betting odds by:

$$p_t^i = \frac{\frac{1}{o_t^i}}{\frac{1}{o_t^i}+\frac{1}{o_t^j}} \tag{1}$$

and

$$p_t^j = \frac{\frac{1}{o_t^j}}{\frac{1}{o_t^i}+\frac{1}{o_t^j}} \tag{2}$$

where O_t^i and O_t^j are the decimal betting odds for a player i win and player j win, respectively, in match t. The denominator in each of these two expressions is the normalization factor (over-round).

The second way expected wins are determined is based on Elo ratings-implied win probabilities. It can be shown that the Bradley-Terry strength/ability parameter can be expressed as a function of the Elo rating (Coulom, 2007). In the Bradley-Terry model, the probability that i beats j is given by:

$$\frac{s_i}{s_i+s_j} \tag{3}$$

where the Elo rating of i is defined as

$$R_i = 400log_{10}(s_i) \tag{4}$$

or equivalently

$$s_i = 10^{\frac{R_i}{400}} \tag{5}$$

Therefore, the Elo-implied estimated win probability for player i over player j in match t is given by:

$$p_t^i = \frac{10^{\frac{R_t^i}{400}}}{10^{\frac{R_t^i}{400}}+10^{\frac{R_t^j}{400}}} \tag{6}$$

Similarly, the Elo-implied estimated win probability for player j over player i in match t is given by:

$$p_t^j = \frac{10^{\frac{R_t^j}{400}}}{10^{\frac{R_t^i}{400}}+10^{\frac{R_t^j}{400}}} \tag{7}$$

Note that, unlike betting odds, normalization is not required as it is already the case with Elo-rating estimated probabilities that $p_t^i + p_t^j = 1$. The before- and after-match Elo ratings were computed by passing the cleaned ATP and WTA datasets (see 2.1. Data) to the `welofit()` function in the welo package (the default options were used). For estimating win probabilities, we use the before-match Elo ratings. Since the betting odds and Elo ratings already account for historical information, it is reasonable to ignore the temporal order of the match result data (i.e., we do not need to consider the match results as a temporally ordered sequence).

2.2.2. Fisher's Exact Test

A Fisher's Exact Test is applied to compare each pair of players' expected wins distribution with their actual match result distribution. The bogey effect is assumed to represent a violation of expectation. The obtained expected wins distribution for each player in each player pair comprises one part of the contingency table to which the Fisher's Exact Test, which has the advantage of being able to be used for small sample sizes, is applied. The expected wins for each player in a given player pair are obtained by simply summing the estimated win probabilities across all matches they have played against that specific player, which represents the expected number of matches each player should win given the betting odds/Elo ratings distribution. The other part of the contingency table is the actual win distribution between the two players. The null hypothesis, H0, and alternative hypothesis, H1, are described as follows:

H_0: The expected and actual win distributions are the same, or there is no significant difference between them.
H_1: The expected and actual win distributions are not the same, or there is a significant difference between them.

The contingency table for the Fisher exact test for one pair of players, player i and player j, is shown in Table 1. In Table 1, N represents the total number of historical matches played between player i and player j. Thus, $\sum_{t=1}^{N} p_t^i$ represents the aggregation of the betting odds- or Elo-implied win probabilities for player i over all historical matches against player j. Similarly, $\sum_{t=1}^{N} p_t^j$ represents the aggregation of the betting odds (or Elo) implied win probabilities for player j over all historical matches against player i. The cell a_t^i in Table 1 is a binary variable that takes the value of 1 if player i won against player j in match t and 0 otherwise, and analogously, a_t^j is a binary variable that takes the value of 1 if player j beat player i in match t and 0 otherwise. Therefore, $\sum_{t=1}^{N} a_t^i$ represents the total number of actual wins player i has had over player j in their N past matches and $\sum_{t=1}^{N} a_t^j$ the total number of actual wins player j has had over player i in their N past matches.

Table 1: Contingency table for Fisher's Exact Test for a given player pair over their *N* historical matches.

	Betting Odds- or Elo ratings-implied win probabilities	Actual match result
Player i wins	$\sum_{t=1}^{N} p_t^i$	$\sum_{t=1}^{N} a_t^i$
Player j wins	$\sum_{t=1}^{N} p_t^j$	$\sum_{t=1}^{N} a_t^j$

2.3. Bogey effect identification

To identify the bogey effect between a pair of players, the method iteratively computes the Fisher's Exact Test *p*-values for each player pair. Having a $p \leq \alpha$, where α is a specific significance level (in this study we consider $\alpha = 0.05$ and $\alpha = 0.1$, which means that we have 95% or 90% confidence that a bogey player pair exists), is a necessary but not sufficient condition for the bogey effect to exist. The *p*-value is calculated from the Fisher's Exact Test that is applied to the contingency table structure for the *N* historical matches between a particular player pair as per Table 1. If Fisher's Exact Test yields a statistically significant result *and* the expected wins and actual wins contradict—that is, player A was expected to win more matches than player B but actually player A won fewer, or player B was expected to win more matches than player A but actually player B won fewer—we suggest that the bogey effect exists between the two players. Using the notation in Table 1, we suggest that the bogey effect exists between a pair of players player i and player j, based on their N past matches, if the following holds true:

$$\begin{aligned} &\text{IF} \quad p \leq \alpha \\ &\text{AND} \quad [(\textstyle\sum_{t=1}^{N} p_t^i > \sum_{t=1}^{N} p_t^j \text{ AND } \sum_{t=1}^{N} a_t^j > \sum_{t=1}^{N} a_t^i) \\ &\text{OR} \quad (\textstyle\sum_{t=1}^{N} p_t^j > \sum_{t=1}^{N} p_t^i \text{ AND } \sum_{t=1}^{N} a_t^i > \sum_{t=1}^{N} a_t^j)] \end{aligned}$$

3. Results

3.1. Number of bogey player pairs identified with the proposed method using Elo ratings and betting odds for each dataset

A summary of the results for all datasets, at the 90% and 95% level of confidence, is shown in Table 2. An initial observation from Table 2 is that, regardless of the level of significance used and whether betting odds or Elo ratings are used, the number of bogey player pairs identified is very small relative to the number of player pairs in the datasets. This suggests that the bogey player effect is very rare in professional tennis.

For the whole ATP dataset, of the 18,241 distinct ATP player pairs, 4 and 18 significant bogey pairs were identified at the 90% significance level using betting odds and Elo ratings, respectively. For the whole WTA dataset, of the 15,844 distinct WTA player pairs, 7 and 13 significant bogey pairs were identified using betting odds and Elo ratings, respectively, at the 90% level of significance.

Table 2: Summary of the number of significant player pairs with the bogey effect identified for each of the six datasets, using aggregated betting odds- and Elo ratings-implied probabilities for computing expected wins, at the 95% and 90% significance levels.

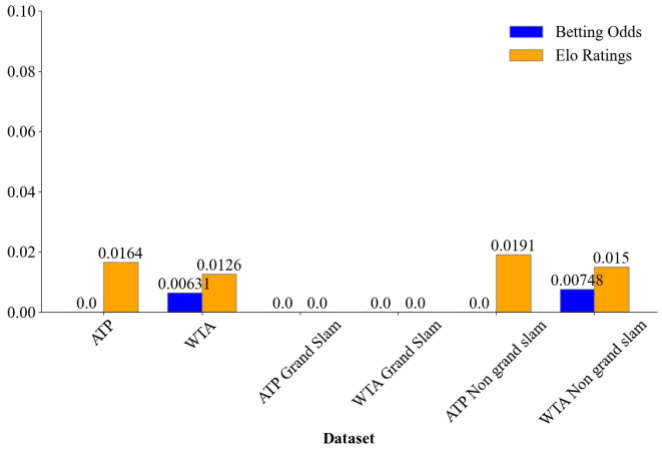
		95% level of significance		90% level of significance	
	Player pairs (*n*)	Betting Odds	Elo Ratings	Betting Odds	Elo Ratings
All					
ATP	18,241	0	3	4	18
WTA	15,844	1	2	7	13
Grand Slam					
ATP	5,485	0	0	1	0
WTA	5,075	0	0	0	0
Non-Grand Slam					
ATP	15,735	0	3	6	12
WTA	13,372	1	2	5	8

Notes: The player pairs that are significant at the 95% significance level are also significant at the 90% significance level. For example, for ATP data with Elo ratings, the 18 significant bogey player pairs at the 90% level of significance also include the 3 player pairs that are significant at the 95% level.

Since the total number of player pairs differs across datasets (see the third column in Table 2), in order to make a like-for-like comparison between datasets, the number of bogey player pairs identified is scaled by the total number of player pairs in the dataset in Figure 5. In particular, Figure 5 shows the number of bogey effect player pairs identified for each dataset using betting odds- and Elo ratings-implied probabilities, as a percentage of the total number of player pairs in each dataset, at the 95% (Figure 5A) and 90% (Figure 5B) significance levels.

Some observations can be made from Figure 5 and Table 2. The proposed method with betting odds-implied probabilities obtained fewer bogey player pairs (as a percentage of total player pairs) for all datasets apart from the ATP Grand Slam match dataset. Across the ATP Grand Slam and WTA Grand Slam datasets with both Elo ratings and betting odds-implied probabilities used to compute expected wins, and the ATP Grand Slams dataset with betting odds-implied probabilities used, only one bogey player pair was identified. These results lend support to what was expected: that Grand Slams are closely followed by bookmakers and thus odds are set accurately, and the small differences in Elo ratings between matches among generally top-ranked players at Grand Slam tournaments result in fewer bogey player pairs being identified compared to in the non-Grand Slam match dataset. While the proposed method using betting odds identified relatively fewer bogey player pairs in men's than women's tennis in the all-match datasets, this wasn't the case with the Grand Slam or non-Grand Slam subsets. Somewhat consistent with Vaughan Williams et al. (2021), Elo ratings appear to perform well in the WTA since there were relatively fewer bogey player pairs identified in the WTA using Elo ratings compared to the ATP.

(A)



(B)

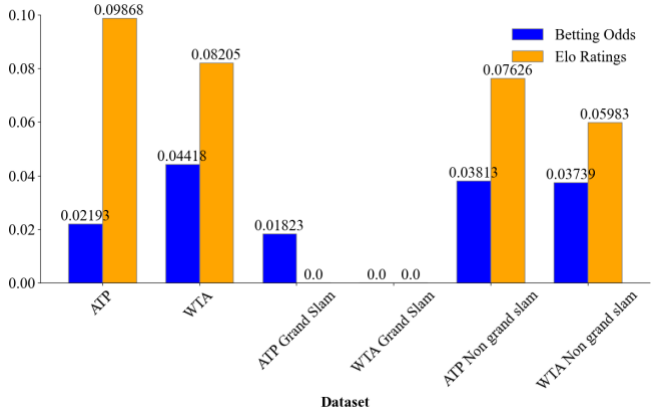


Figure 5: The number of bogey effect player pairs identified for each dataset using betting odds- and Elo ratings-implied probabilities, as a percentage of the total number of player pairs in each dataset, at the 95% (A) and 90% (B) significance levels (data from Table 2).

3.2. Visualising obtained bogey player pairs for a particular dataset

Since the order of the players in each player pair is not material, we visualise bogey player pairs on the absolute difference in actual wins against the absolute difference in expected wins. The size of these points is scaled based on the FET *p*-value, and significant player pairs based on the FET that are not bogey player pairs can be distinguished by shape. Figures 6 to Figure 9 depict the significant bogey (and significant non-bogey) player pairs in this manner for the all-match ATP and WTA datasets, obtained with Elo ratings- and betting odds-implied probabilities (Supplementary Figures 5 to 8/Supplementary Tables 6 to 9 show the same plots and corresponding data for the Grand Slam and non-Grand Slam datasets for the ATP and WTA). These figures correspond to the data in Supplementary Tables 1, 2, 3, and 4 in the Appendix.

In general, Figures 6 to Figure 9 exhibit a cluster of points towards the bottom-left hand corner with an absolute actual win difference between the players in the bogey player pair of around two and an absolute expected wins difference of around three. There is often an additional cluster of points with roughly the same absolute expected wins difference but higher absolute actual wins difference values. In all figures there was also an outlier player pair with higher absolute actual and expected wins differences, however, these were generally significant but not identified as a bogey player pair since their actual and expected wins did not contradict.

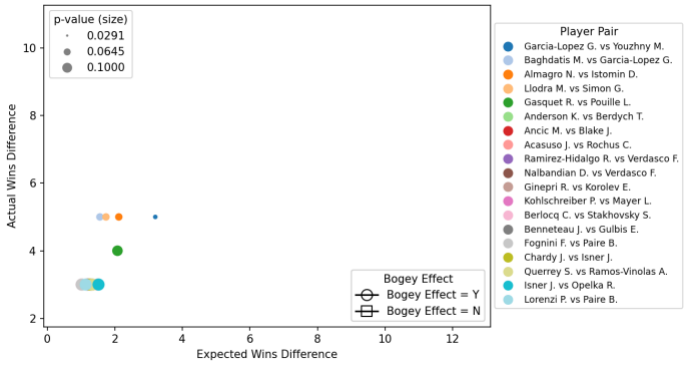


Figure 6: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, the statistically significant (with at least 90% confidence) and bogey player pairs from the ATP tennis dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins. The data used to create this plot is presented in Supplementary Table 1 in the Appendix.

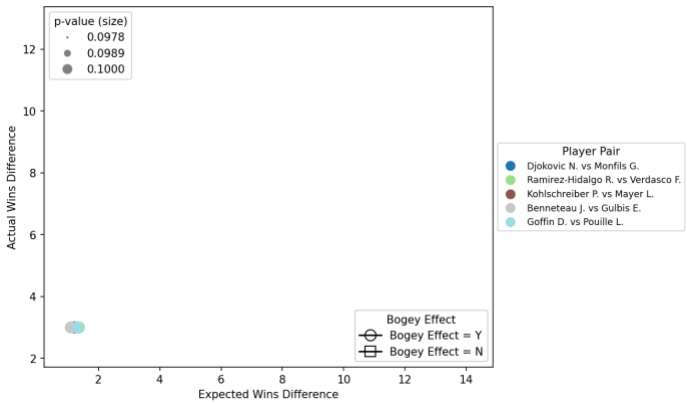


Figure 7: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) and bogey player pairs from the ATP tennis dataset, with aggregated betting odds-implied probabilities used to compute expected wins. The data used to create this plot is presented in Supplementary Table 2 in the Appendix.

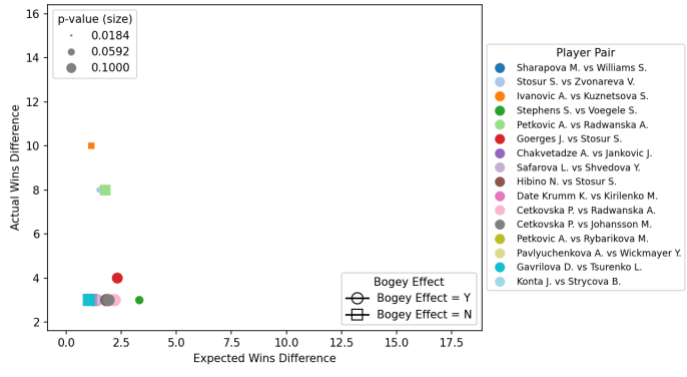


Figure 8: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically

significant (with at least 90% confidence) bogey player pairs from the WTA tennis dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins. The data used to create this plot is presented in Supplementary Table 3 in the Appendix.

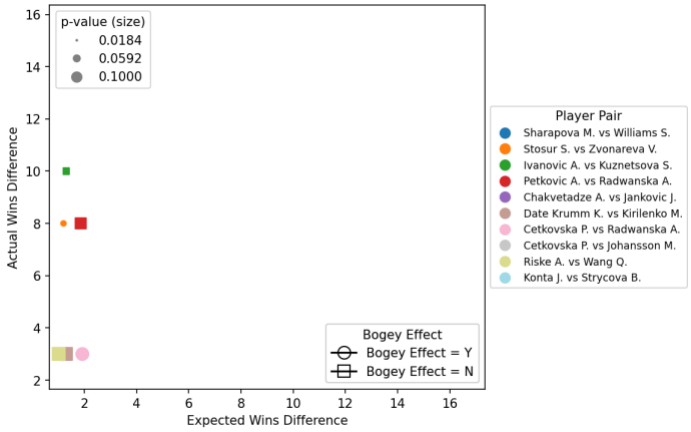
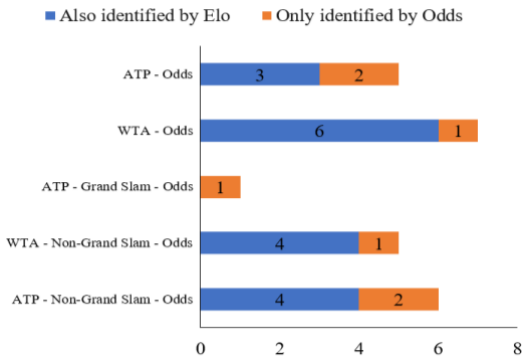


Figure 9: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) bogey player pairs from the WTA tennis dataset, with aggregated Betting odds-implied probabilities used to compute expected wins. The data used to create this plot is presented in Supplementary Table 4 in the Appendix.

3.3. Analysing the overlap in the bogey player pairs identified by Elo ratings and betting odds

Figure 10A shows the number of bogey player pairs identified by betting odds based on whether they were also identified by Elo ratings (or only betting odds). Figure 10B shows the number of bogey player pairs identified by Elo ratings based on whether the player pairs were also identified by betting odds (or only by Elo ratings).

(A)



(B)

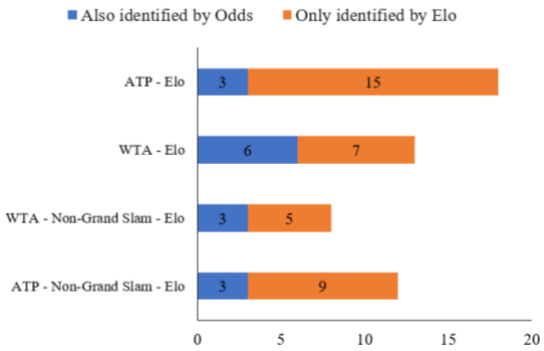


Figure 10: The number of bogey player pairs identified by (A) betting odds, split into whether they were also identified by Elo ratings or only by betting odds, and (B) Elo ratings, split into whether they were also identified by betting odds or only Elo ratings.

For example, for the whole ATP dataset, of the significant bogey player pairs identified by Elo ratings (Supplementary Table 1), three of these (Ramirez-Hidalgo R. vs Verdasco F.; Kohlschreiber P. vs Mayer L.; and Benneteau J. vs Gulbis E.) were also identified by betting odds (Supplementary Table 2). Fifteen bogey player pairs that were identified by Elo ratings were only identified by Elo but were not identified by betting odds.

Figure 10A suggests that the bogey player pairs identified by the proposed method with betting odds were largely also identified with Elo ratings. However, Figure 10B suggests that the bogey player pairs identified by the proposed method with Elo ratings were largely not also identified with betting odds (the WTA all-match dataset is one notable exception).

3.4. Visualising the expected win distribution violation quantification for each dataset

Figure 11 shows average expected and actual win probability, as well as the differences, for player pairs containing and not containing bogey players. The underlying data for this plot is shown in Supplementary Table 12 in the Appendix. For each player pair type, whether the player pair is a bogey player pair or not, the average expected wins and average actual wins were calculated by scaling the actual and expected wins based on the total number of matches between the players in the player pair. Taking the difference between these two values provides a means of quantifying the degree to which the expected win distribution is violated. The red values in Figure 11 denote the average difference in expected and actual win probabilities for player pairs without a bogey player, while the blue values denote the average difference in expected and actual win probabilities for player pairs that don't involve a bogey player. Three of the datasets/methods on the far-right have no bogey player pairs (see Table 2) and therefore only two points show for these. As we would expect based on our proposed method's design, for player pairs that do contain a bogey player, there is a large violation of the expected wins distribution in terms of the average difference in expected and actual wins (the blue values), and are many times larger than the violation of the expected distribution for player pairs not containing a bogey player (red values). It is also notable that while the values representing the violation of the expected wins distribution for bogey player pairs were all relatively similar, ranging from 0.665 to 0.708, whereas the values representing the violation of the expected wins distribution

values for bogey player pairs—while smaller in magnitude—ranged from 0.0389 to -0.0382.

4. Discussion

This study proposed a bogey player identification method that involves computing an expected wins distribution using the summed implied win probabilities based on Elo ratings and betting odds, constructing a contingency table containing actual and expected wins between each player in a player pair, and applying Fisher's Exact Test to this contingency table. If a significant result from Fisher's Exact Test was obtained, and the actual wins and expected wins contradict, the bogey effect was deemed to exist between the two players in the player pair.

The obtained results suggest that although the bogey player effect exists in professional tennis, it is very rare. The proposed

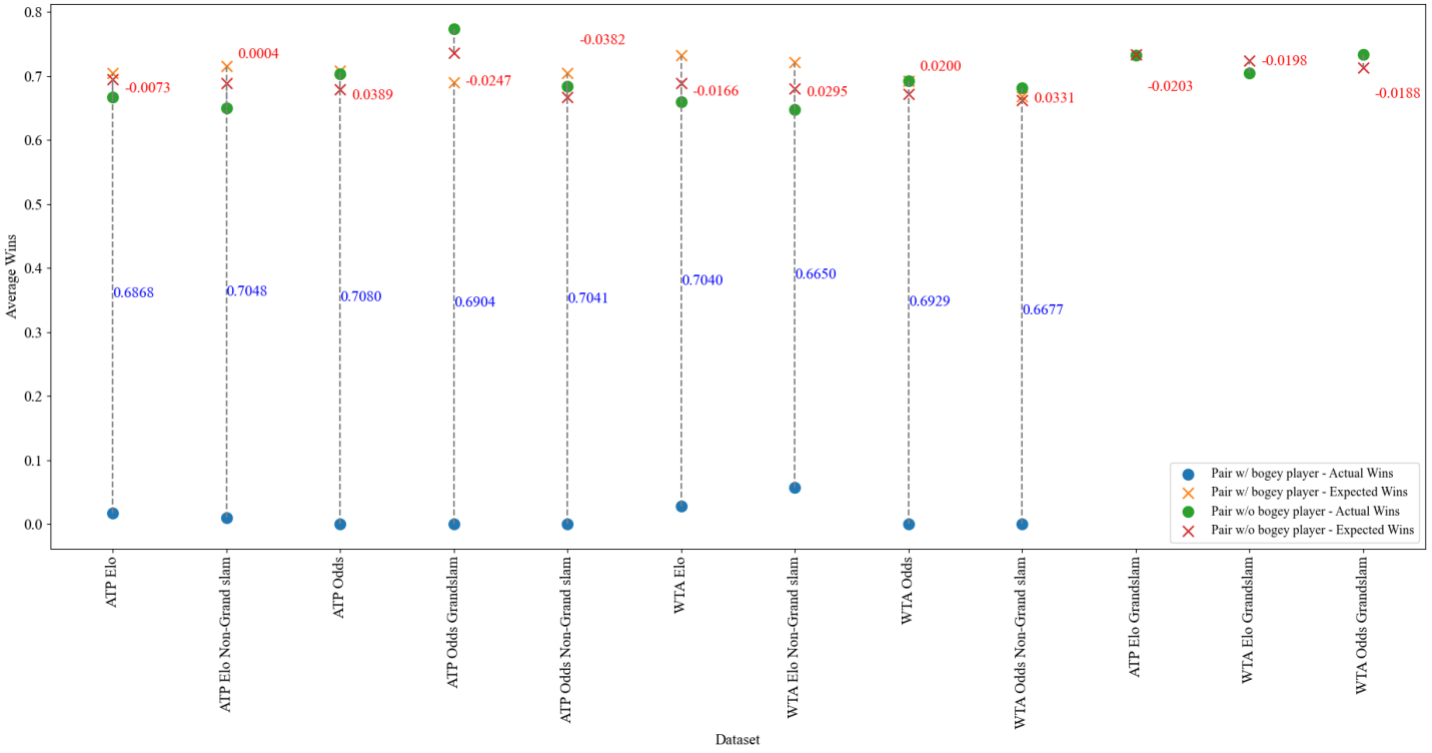


Figure 11: For each dataset and method (Elo/odds), this plot shows the average expected wins and average actual wins scaled based on the number of matches between players in a player pair, as well as their differences for player pairs that contain and do not contain bogey players (at the 90% level of confidence) (see Supplementary Table 12 in the Appendix for the underlying data in this plot).

method with betting odds used to compute expected wins obtained fewer bogey player pairs, as a percentage of total player pairs in the dataset, for all datasets except for the ATP Grand Slam dataset. Only one significant (at the 90% level of confidence) bogey player pair was identified across the four Grand Slam datasets: ATP and WTA Grand Slam datasets using Elo ratings and betting odds. The results conformed with our prior expectation that Grand Slams are closely analysed by bookmakers and so odds are set accurately, which results in fewer bogey player pairs in general. Furthermore, when the proposed method is used with Elo ratings to compute expected wins, the small differences in Elo ratings in matches among generally top-ranked players at Grand Slams mean that fewer bogey player pairs are identified compared to non-Grand Slams. Although betting odds identified relatively fewer bogey player pairs in ATP than WTA tennis in the ATP and WTA datasets as a whole, this did not hold for the Grand Slam or non-Grand Slam datasets. Since there were relatively fewer bogey player pairs identified in the WTA using Elo ratings compared to the ATP, Elo ratings could be said to be of better predictive value for the WTA than the ATP, a result that is consistent with Vaughan Williams et al. (2021). Surprisingly, when visualising obtained FET-significant player pairs and bogey player pairs for the various datasets, certain patterns emerged in terms of the clusters of bogey player pairs' absolute differences in actual and expected wins. However, these patterns/clusters did not appear useful for identifying which of the pairs are actually bogey player pairs. When analysing the overlap in the bogey player pairs that were identified by Elo ratings and betting odds, while betting odds generally identified fewer bogey player pairs than Elo ratings, the majority of the bogey player pairs identified by betting odds were also identified by Elo ratings. On the other hand, the majority of bogey player pairs identified by Elo ratings were only identified by Elo ratings but not by betting odds. This suggests perhaps that betting odds may be generally a more reliable means of computing expected wins and thus identifying bogey player pairs. When visualising the expected win distribution violation quantification for the various datasets by considering the average difference in expected and actual wins for player pairs containing and not containing bogey players, we validated that, for bogey player pairs, there was a large violation of the expected wins distribution in terms of the average difference in expected and actual wins, and these differences – which were relatively similar across the

different datasets/methods (Elo and odds) – were many times larger than the violation of the expected distribution for non-bogey player pairs.

Analysing a particular player's performance, whether they are prone to being a bogey player or being the bogey player of another, can be useful for player-level performance analysis and match preparation. For instance, Stosur is a WTA player who appeared both as a bogey player and as a non-bogey player in bogey player pairs, which is interesting from a practical performance analysis perspective, for example, for her opponents and coaching staff to analyse further.

The proposed method is flexible in that it can be applied not only to tennis but to other sports, directly to sports with two outcomes and with some modifications to sports with more than two outcomes. In future work, the method could therefore be applied to other sports with two outcomes (e.g., basketball), and it could be extended to sports with three outcomes, for example, soccer, by using extensions of Fisher's Exact Test that can handle contingency tables with more than two columns/rows, for example, the Freeman-Halton extension of Fisher's Exact Test (Freeman & Halton, 1951). For instance, to include a draw outcome, in addition to summing the win probability for each opposition as in the current study, the probability of a draw multiplied by 0.5 for both players would be summed, and 0.5 could be used to represent an actual draw result. Rating systems other than Elo ratings could also be trialled. Finally, other subsets of the original dataset other than Grand Slam/non-Grand Slam, for example, based on court surface, time period, player hand, and player rank group could be considered. For example, based on Figure 3 above, splitting the WTA dataset into 2005 to 2012 and 2013 to 2020 would be an obvious partition of the original dataset. Also, subsets of the original dataset based on Elo rating differences would also be interesting to investigate. For instance, since bookmakers have only a limited (or no match) history to go on in the case of matches among two low-ranked players, odds may be more difficult to set, and therefore their value for prediction may be lower and thus more bogey player pairs would be identified. Matches among low-ranked players would, however, have a small ranking difference in terms of Elo ratings and, therefore, based on the findings of Yue et al. (2022) would be more predictable and we could hypothesise that this would result in fewer bogey player pairs compared to when odds are used.

Conflict of Interest

The authors declare no conflict of interests.

Acknowledgment

This work was supported by JSPS under Grant [number 20H04075] and JST Presto under Grant [number JPMJPR20CA].

Data availability

The dataset that supports the findings of this study was obtained, as described in the "Data" subsection of "Materials and Methods", from the openly available data in the appendix of the online version of the paper by Angelini, Candila, and De Angelis (2022), at doi:10.1016/j.ejor.2021.04.011 under "Supplementary Data S1". This is open data under the CC BY license http://creativecommons.org/licenses/by/4.0/

Code

The code is available at the following GitHub repository: https://github.com/rorybunker/bogey-phenomenon-sport/

References

Angelini, G., Candila, V., & De Angelis, L. (2022). Weighted Elo rating for tennis match predictions. *European Journal of Operational Research, 297*(1), 120–132.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition, 32*(6), 1369–1378.

Awerbuch, W. (2009). *Anwendung von Data Mining zu statistischen Auswertungen und Vorhersagen im Sport* [Master's thesis, TU Darmstadt]. TU Darmstadt Knowledge Engineering website. https://ke-tud.github.io/lehre/arbeiten/diplom/2009/Awerbuch_Wladimir.pdf

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting, 35*(2), 741–755.

Barrutiabengoa, J. M., Corredor, P., & Muga, L. (2022). Does the betting industry price gender? Evidence from professional tennis. *Journal of Sports Economics, 23*(7), 881–906.

Bar-Eli, M., Avugos, S., & Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise, 7*(6), 525–553.

Bunker, R. (2022). The Bogey Phenomenon in Sport. In J. J. Reade (Eds.), *IX Mathsport International 2022 Proceedings* (pp.15–21). https://www.mathsportinternational.com/MathSport2022Proceedings.pdf

Candila, V. (2023). welo: An R package for weighted and standard Elo rates. *Statistica Applicata-Italian Journal of Applied Statistics*, (1).

Carlson, K. A., & Shu, S. B. (2007). The rule of three: How the third event signals the emergence of a streak. *Organizational Behavior and Human Decision Processes, 104*(1), 113–121.

Chiweshe, M. K. (2021). Frenemies: Understanding the interconnectedness of rival fan identities in Harare, Zimbabwe. In K. Bandyopadhyay (Eds.), *Face to Face: Enduring Rivalries in World Soccer* (pp. 191–203). Routledge.

Cattelan, M., Varin, C., & Firth, D. (2013). Dynamic Bradley–Terry modelling of sports tournaments. *Journal of the Royal Statistical Society Series C: Applied Statistics, 62*(1), 135–150.

Coulom, R. (2007). Computing "elo ratings" of move patterns in the game of go. *ICGA Journal, 30*(4), 198–208.

Estes, W. K. (1964). Probability learning. In A. Melton (Ed.), *Categories of human learning* (pp. 89–128). Academic Press/Elsevier.

Freeman, G. H., & Halton, J. H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika, 38*(1/2), 141–149.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*(3), 295–314.

Hales, S. D. (1999). An epistemologist looks at the hot hand in sports. *Journal of the Philosophy of Sport, 26*(1), 79–87.

Hankin, R. K. S., & Bunker, R. (2016). Bogey teams in sport. Technical report, Auckland University of Technology.

Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports, 12*(3), 127–138.

Leitner, C., Zeileis, A., & Hornik, K. (2009). Is Federer stronger in a tournament without Nadal? An evaluation of odds and seedings for Wimbledon 2009. *Austrian Journal of Statistics, 38*(4), 277–286.

Niall, J. (2013). It's Murray v Djokovic. The Sydney Morning *Herald.* https://www.smh.com.au/sport/tennis/its-murray-v-djokovic-20130125-2dcuv.html

Peel, L., & Clauset, A. (2015). Predicting sports scoring dynamics with restoration and anti-persistence. In C. Aggarwal, Z. Zhou, A. Tuzhilin, H. Xiong, & X. Wu (Eds.), *2015 IEEE International Conference on Data Mining* (pp. 339–348). IEEE. https://doi.org/10.1109/ICDM36327.2015

Poulton, E. (2004). Mediated patriot games: The construction and representation of national identities in the British television production of Euro'96. *International Review for the Sociology of Sport, 39*(4), 437–455.

Raab, M. (2012). Simple heuristics in sports. *International Review of Sport and Exercise Psychology, 5*(2), 104–120.

Steeger, G. M., Dulin, J. L., & Gonzalez, G. O. (2021). Winning and losing streaks in the National Hockey League: Are teams experiencing momentum or are games a sequence of random events? *Journal of Quantitative Analysis in Sports, 17*(3), 155–170.

Stone, D. F. (2012). Measurement error and the hot hand. *The American Statistician, 66(*1), 61–66.

Vaughan Williams, L., Liu, C., Dixon, L., & Gerrard, H. (2021). How well do Elo-based ratings predict professional tennis matches? *Journal of Quantitative Analysis in Sports, 17*(2), 91–105.

Wilms, S. (2014). Mental Training im Sport: Möglichkeiten der Anwendung in der Sozialen Arbeit [Doctoral dissertation, Hochschule für angewandte Wissenschaften Hamburg]. HAW Hamburg Repository. https://reposit.haw-hamburg.de/handle/20.500.12738/6563

Wood, L. (2017). Jo-Wilfried Tsonga hopes to avoid Australian Open bogey Kei Nishikori. *Herald Sun.* https://www.heraldsun.com.au/sport/tennis/jowilfried-tsonga-hopes-to-avoid-australian-open-bogey-kei-nishikori/news-story/2dfaab3f641494447405ae65fc7e7592

Yue, J. C., Chou, E. P., Hsieh, M. H., & Hsiao, L. C. (2022). A study of forecasting tennis matches via the Glicko model. *PLoS ONE, 17*(4), 1–12. https://doi.org/10.1371/journal.pone.0266838

Supplementary materials

Supplementary Table 1: Statistically significant (with at least 90% confidence) and bogey player pairs from the ATP tennis dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Garcia-Lopez G. vs Youzhny M.	0.0291	1.9/5.1	6/1	Y
Baghdatis M. vs Garcia-Lopez G.	0.0476	3.28/1.72	0/5	Y
Almagro N. vs Istomin D.	0.0476	3.56/1.44	0/5	Y
Llodra M. vs Simon G.	0.0476	1.63/3.37	5/0	Y
Gasquet R. vs Pouille L.	0.0801	4.04/1.96	1/5	Y
Anderson K. vs Berdych T.	0.0932	3.07/8.93	0/12	N
Ancic M. vs Blake J.	0.1	0.91/2.09	3/0	Y
Acasuso J. vs Rochus C.	0.1	2.25/0.75	0/3	Y
Ramirez-Hidalgo R. vs Verdasco F.	0.1	0.85/2.15	3/0	Y
Nalbandian D. vs Verdasco F.	0.1	2.1/0.9	0/3	Y
Ginepri R. vs Korolev E.	0.1	2.16/0.84	0/3	Y
Kohlschreiber P. vs Mayer L.	0.1	2.15/0.85	0/3	Y
Berlocq C. vs Stakhovsky S.	0.1	2.01/0.99	0/3	Y
Benneteau J. vs Gulbis E.	0.1	0.85/2.15	3/0	Y
Fognini F. vs Paire B.	0.1	2.01/0.99	0/3	Y
Chardy J. vs Isner J.	0.1	0.87/2.13	3/0	Y
Querrey S. vs Ramos-Vinolas A.	0.1	2.17/0.83	0/3	Y
Isner J. vs Opelka R.	0.1	2.26/0.74	0/3	Y
Lorenzi P. vs Paire B.	0.1	0.93/2.07	3/0	Y

Notes: As mentioned in the manuscript when describing the proposed method, the FET needs to have a statistically significant *p*-value but also the expected wins and actual wins need to contradict to be suggestive of the bogey effect between a particular player pair. In Supplementary Table 1 above, there is one case where the expected wins and actual wins did not contradict. In particular, the FET *p*-value for Anderson vs Berdych was significant with 1- $\alpha = 1 - 0.1 = 90\%$ confidence (*p*-value = $0.0932 \leq \alpha$), however, the expected wins and actual wins do not contradict. Out of the $N = 12$ historical matches between Anderson and Berdych in the dataset, Anderson was expected, based on the sums of the respective players' Elo ratings-implied win probabilities, to win 3.07 of the matches, while Berdych was expected to win 8.93 of the matches. Berdych did better than expected, achieving 12 wins while Anderson achieved zero. Thus, although this was a statistically significant result, since the actual wins and expected wins did not contradict, this is not a case where the bogey effect was deemed to exist.

Supplementary Table 2: Statistically significant (with at least 90% confidence) and bogey player pairs from the ATP tennis dataset, with aggregated betting odds-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Djokovic N. vs Monfils G.	0.0978	10.97/3.03	14/0	N
Ramirez-Hidalgo R. vs Verdasco F.	0.1	0.82/2.18	3/0	Y
Kohlschreiber P. vs Mayer L.	0.1	2.11/0.89	0/3	Y
Benneteau J. vs Gulbis E.	0.1	0.95/2.05	3/0	Y
Goffin D. vs Pouille L.	0.1	2.16/0.84	0/3	Y

Note: All player pairs in Table 4 are deemed bogey player pairs apart from Djokovic N. vs Monfils G. since the expected and actual wins do not contradict.

Supplementary Table 3: Statistically significant (with at least 90% confidence) bogey player pairs from the WTA tennis dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Sharapova M. vs Williams S.	0.0184	5.56/11.44	0/17	N
Stosur S. vs Zvonareva V.	0.0256	3.25/4.75	8/0	Y
Ivanovic A. vs Kuznetsova S.	0.0325	5.57/4.43	10/0	N
Stephens S. vs Voegele S.	0.0476	4.16/0.84	1/4	Y
Petkovic A. vs Radwanska A.	0.0769	3.12/4.88	0/8	N
Goerges J. vs Stosur S.	0.0801	1.84/4.16	5/1	Y
Chakvetadze A. vs Jankovic J.	0.1	0.87/2.13	3/0	Y
Safarova L. vs Shvedova Y.	0.1	2.2/0.8	0/3	Y
Hibino N. vs Stosur S.	0.1	0.59/2.41	3/0	Y
Date Krumm K. vs Kirilenko M.	0.1	0.93/2.07	3/0	N
Cetkovska P. vs Radwanska A.	0.1	0.4/2.6	3/0	Y
Cetkovska P. vs Johansson M.	0.1	2.46/0.54	0/3	Y
Petkovic A. vs Rybarikova M.	0.1	2.09/0.91	0/3	Y
Pavlyuchenkova A. vs Wickmayer Y.	0.1	2.09/0.91	0/3	Y
Gavrilova D. vs Tsurenko L.	0.1	2.02/0.98	0/3	N
Konta J. vs Strycova B.	0.1	2.12/0.88	0/3	Y

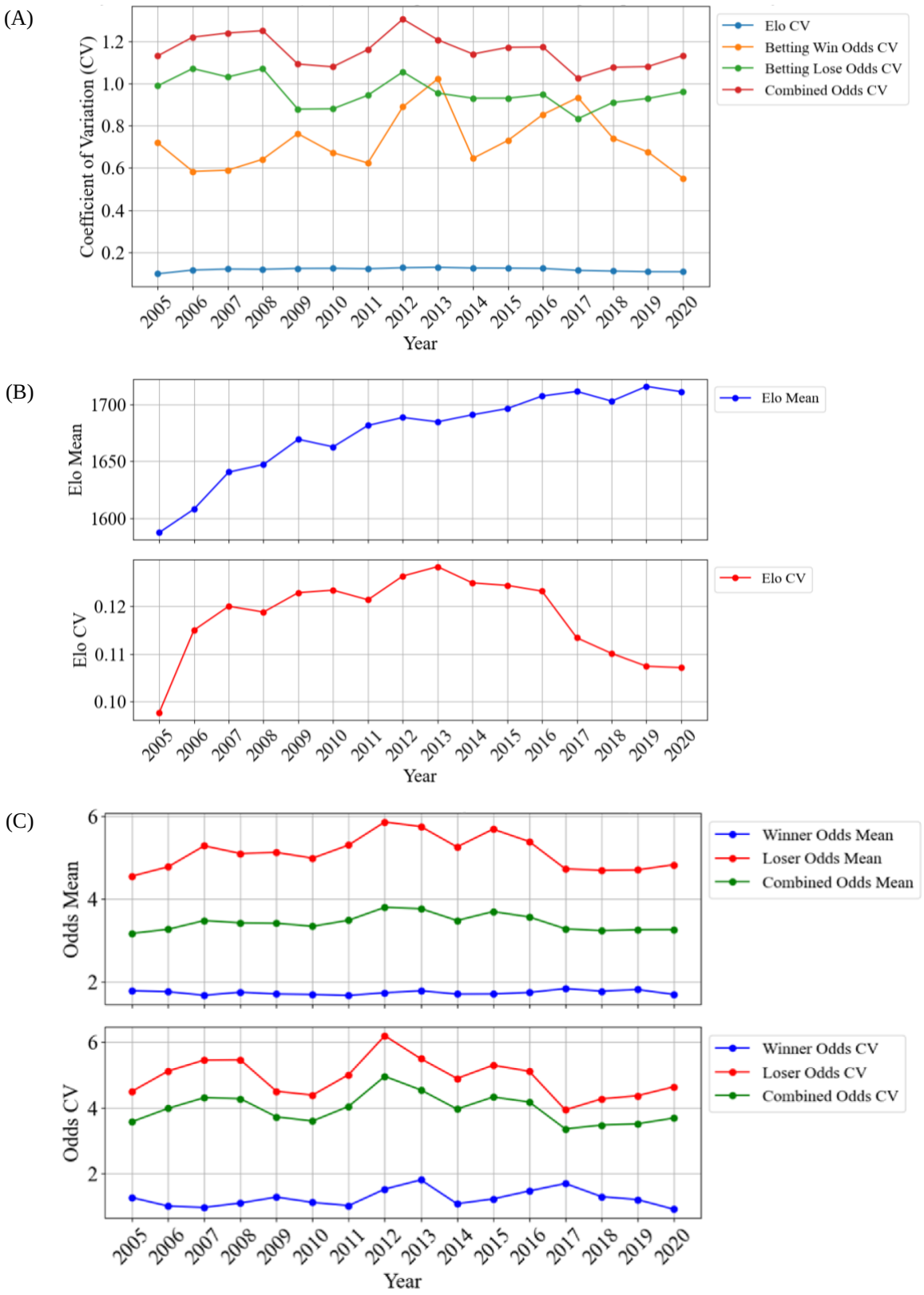
Supplementary Table 4: Statistically significant (with at least 90% confidence) bogey player pairs from the WTA tennis dataset, with aggregated Betting odds-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Sharapova M. vs Williams S.	0.0184	5.4/11.6	0/17	N
Stosur S. vs Zvonareva V.	0.0256	3.4/4.6	8/0	Y
Ivanovic A. vs Kuznetsova S.	0.0325	5.65/4.35	10/0	N
Petkovic A. vs Radwanska A.	0.0769	3.07/4.93	0/8	N
Chakvetadze A. vs Jankovic J.	0.1	0.92/2.08	3/0	Y
Date Krumm K. vs Kirilenko M.	0.1	0.85/2.15	3/0	N
Cetkovska P. vs Radwanska A.	0.1	0.54/2.46	3/0	Y
Cetkovska P. vs Johansson M.	0.1	2.12/0.88	0/3	Y
Riske A. vs Wang Q.	0.1	0.99/2.01	3/0	N
Konta J. vs Strycova B.	0.1	2.02/0.98	0/3	Y

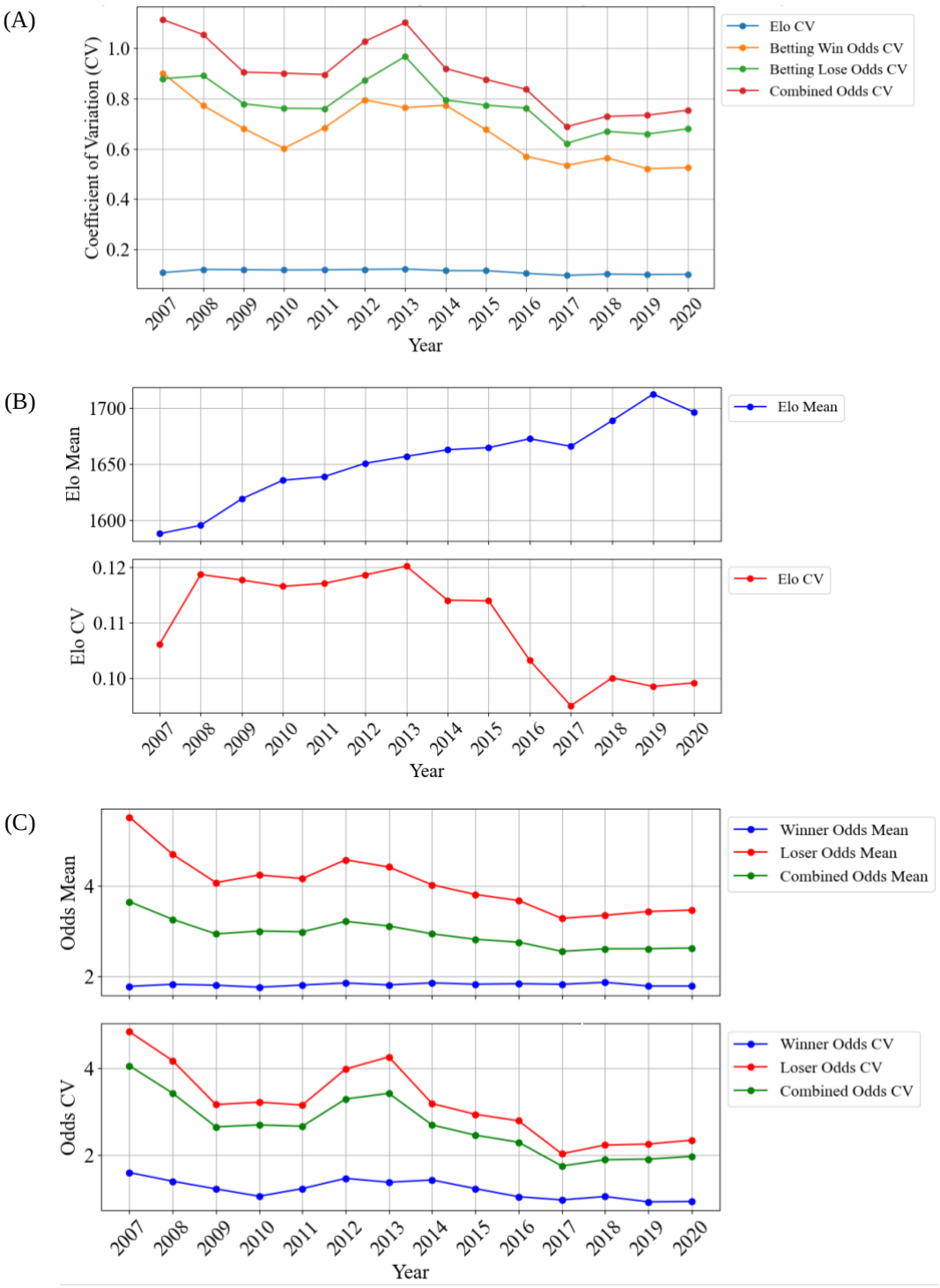
Supplementary Table 5: Statistically significant (with at least 90% confidence) bogey player pairs from the ATP tennis Grand Slams dataset, with aggregated Betting odds-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Isner J. vs Kohlschreiber P.	0.1	2.07/0.93	0/3	Y

Notes: The ATP and WTA Grand Slams datasets with Elo ratings used to compute expected wins, and the WTA Grand Slams dataset with betting odds-implied probabilities used to compute expected wins, all yielded no bogey effect player pairs.

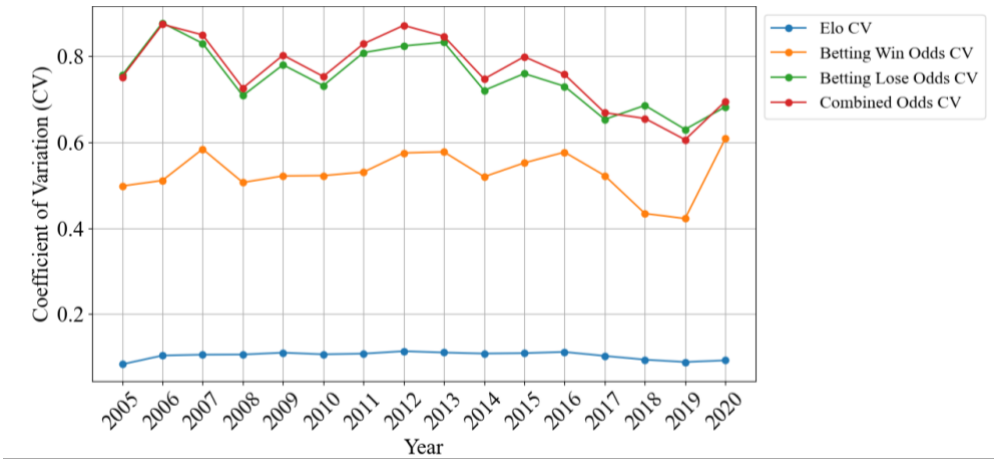


Supplementary Figure 1: Coefficient of Variation (CV) of Betting Odds and Elo Ratings for each year (A), as well as the more granular view of the mean and CV for each year of Elo rating and Betting Odds (B and C, respectively) for the ATP Grand Slam dataset.

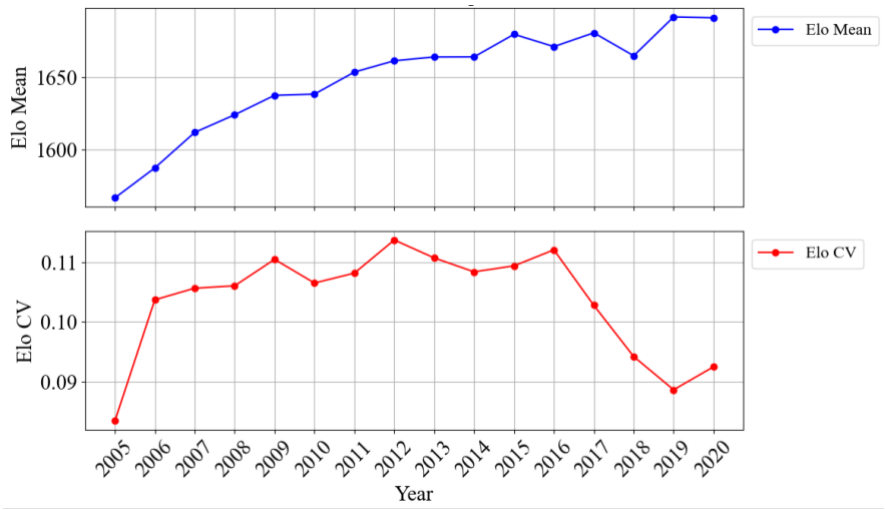


Supplementary Figure 2: Coefficient of Variation (CV) of Betting Odds and Elo Ratings for each year (A) as well as the more granular view of the mean and CV for each year of Elo rating and Betting Odds (B and C, respectively) for the WTA Grand Slam dataset.

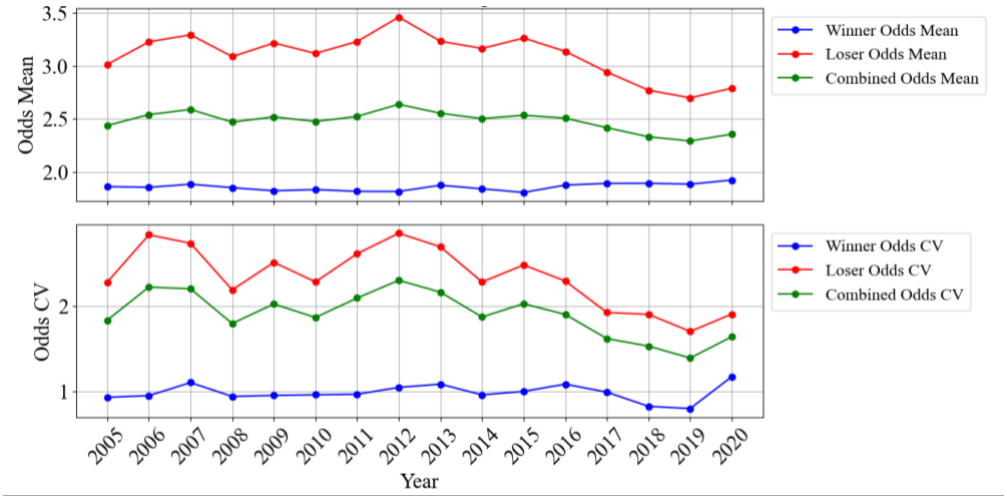
(A)



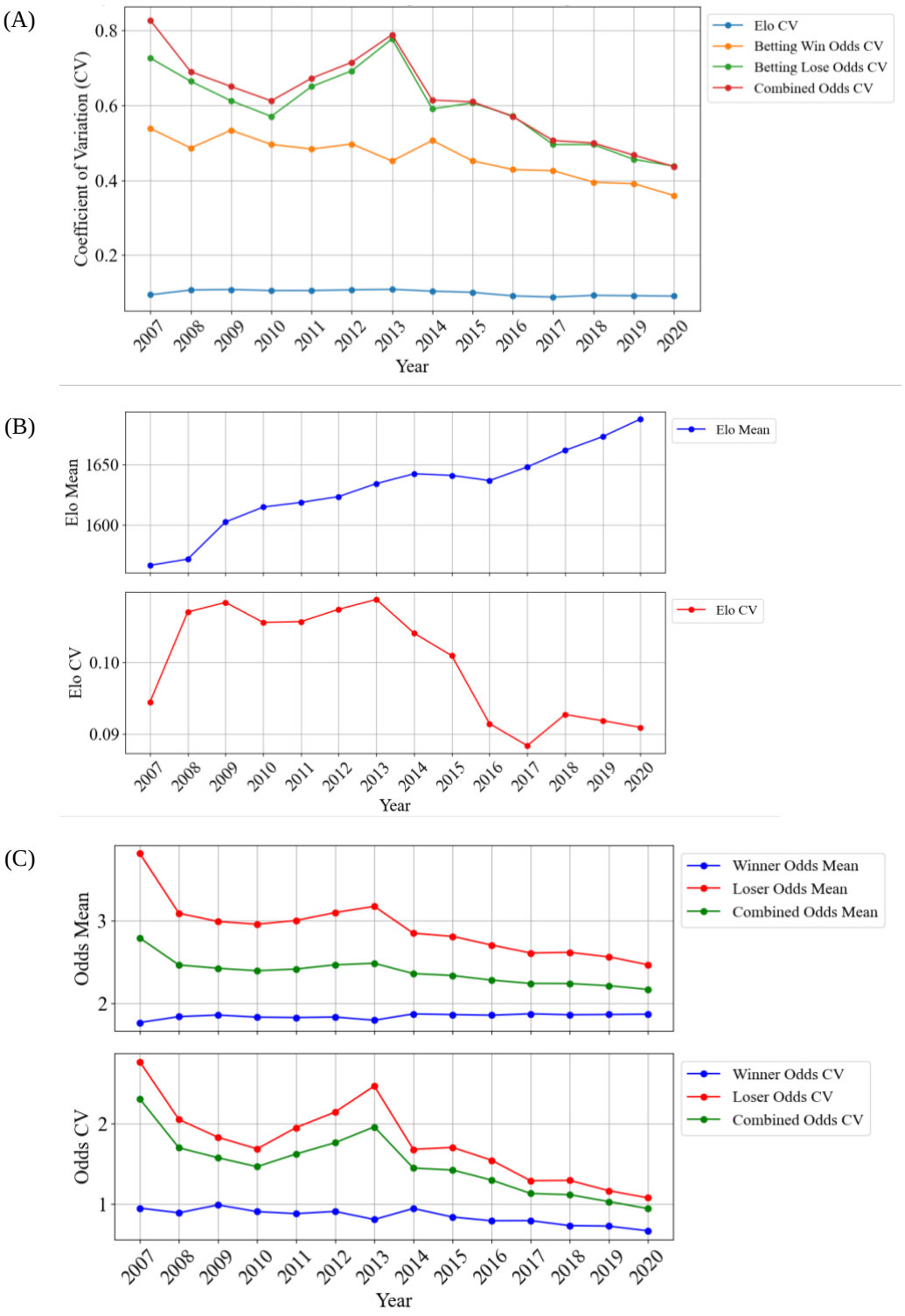
(B)



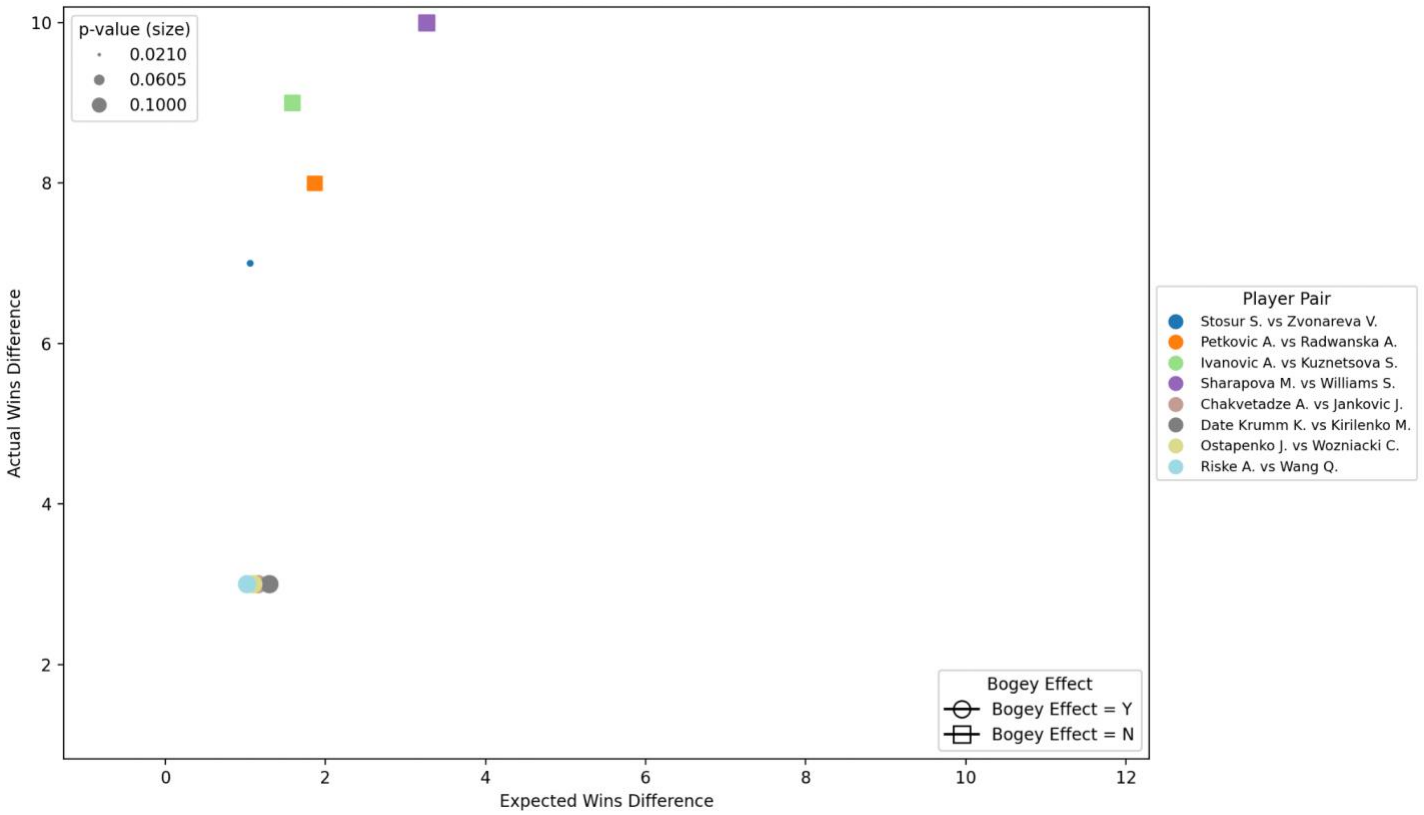
(C)



Supplementary Figure 3: Coefficient of Variation (CV) of Betting Odds and Elo Ratings for each year (A), as well as the more granular view of the mean and CV for each year of Elo rating and Betting Odds (B and C, respectively) for the ATP non-Grand Slam dataset.



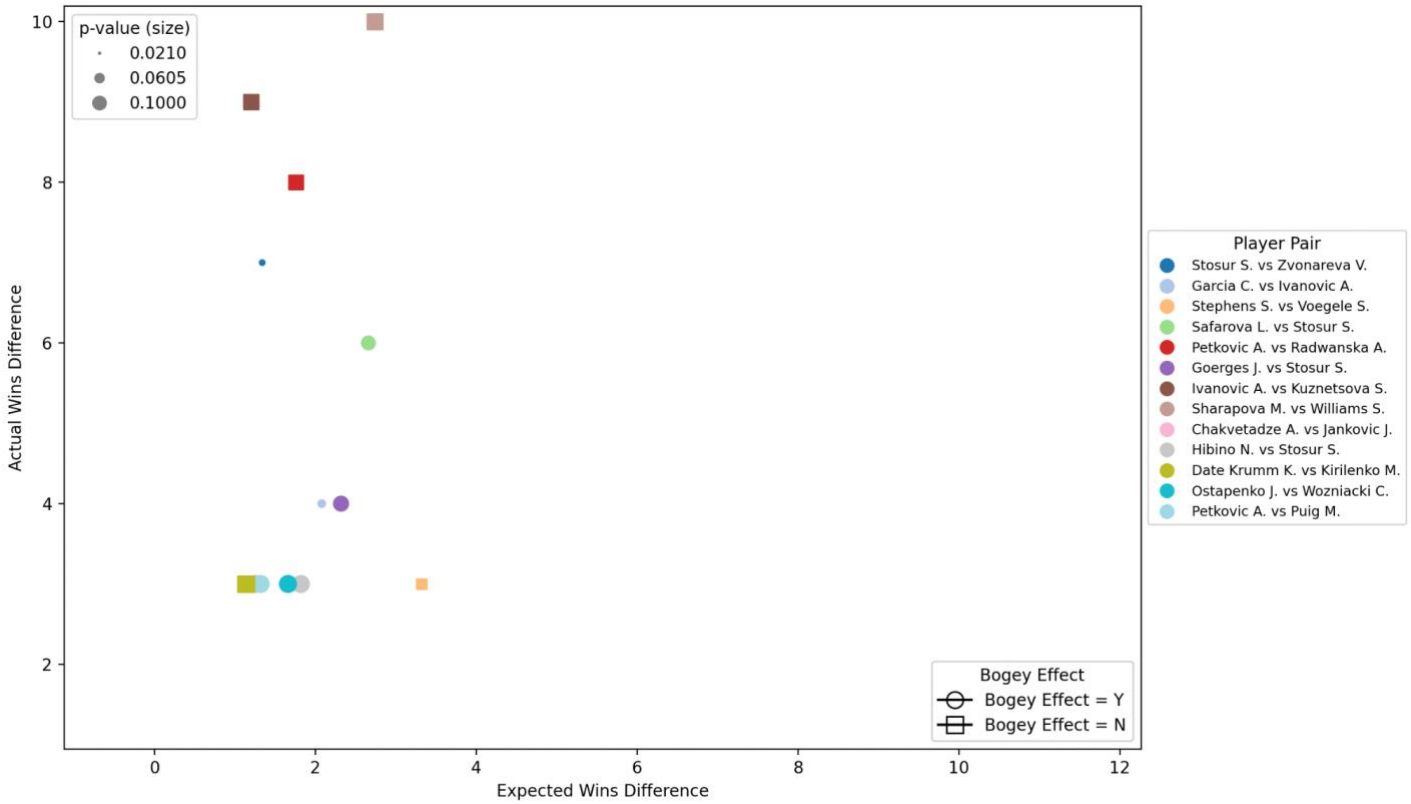
Supplementary Figure 4: Coefficient of Variation (CV) of Betting Odds and Elo Ratings for each year (A), as well as the more granular view of the mean and CV for each year of Elo rating and Betting Odds (B and C, respectively) for the WTA non-Grand Slam dataset.



Supplementary Figure 5: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) bogey player pairs from the WTA tennis non-Grand Slams dataset, with aggregated betting odds-implied probabilities used to compute expected wins. The data used to generate this plot is presented below in Supplementary Table 6.

Supplementary Table 6: Statistically significant (with at least 90% confidence) bogey player pairs from the WTA tennis non-Grand Slams dataset, with aggregated betting odds-implied probabilities used to compute expected wins.

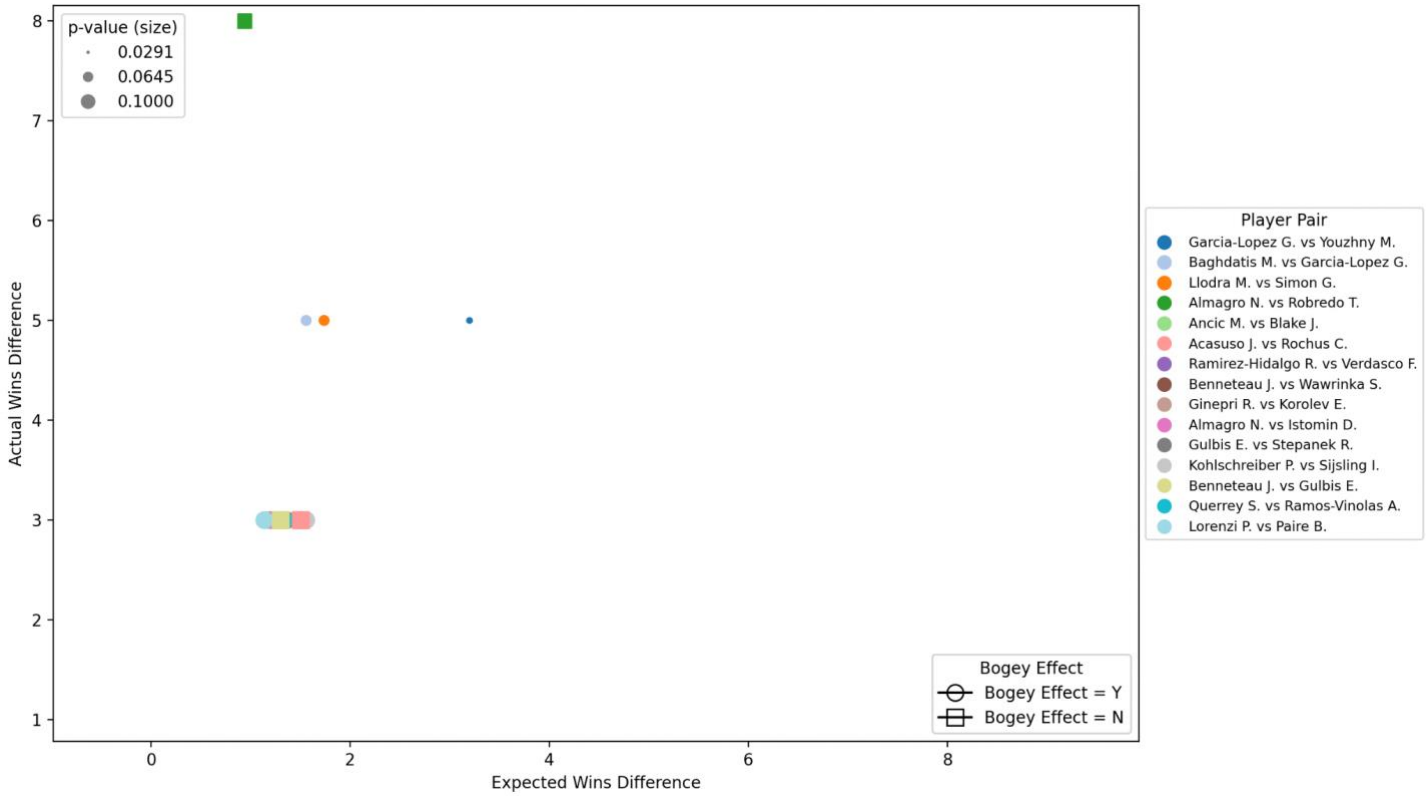
Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Stosur S. vs Zvonareva V.	0.021	2.97/4.03	7/0	Y
Petkovic A. vs Radwanska A.	0.0769	3.07/4.93	0/8	
Ivanovic A. vs Kuznetsova S.	0.0824	5.29/3.71	9/0	
Sharapova M. vs Williams S.	0.0867	3.37/6.63	0/10	
Chakvetadze A. vs Jankovic J.	0.1	0.92/2.08	3/0	Y
Date Krumm K. vs Kirilenko M.	0.1	0.85/2.15	3/0	Y
Ostapenko J. vs Wozniacki C.	0.1	0.95/2.05	3/0	Y
Riske A. vs Wang Q.	0.1	0.99/2.01	3/0	Y



Supplementary Figure 6: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) bogey player pairs from the WTA non-Grand Slams dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins. The data used to generate this plot is presented below in Supplementary Table 7.

Supplementary Table 7: Statistically significant (with at least 90% confidence) bogey player pairs from the WTA non-Grand Slams dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins.

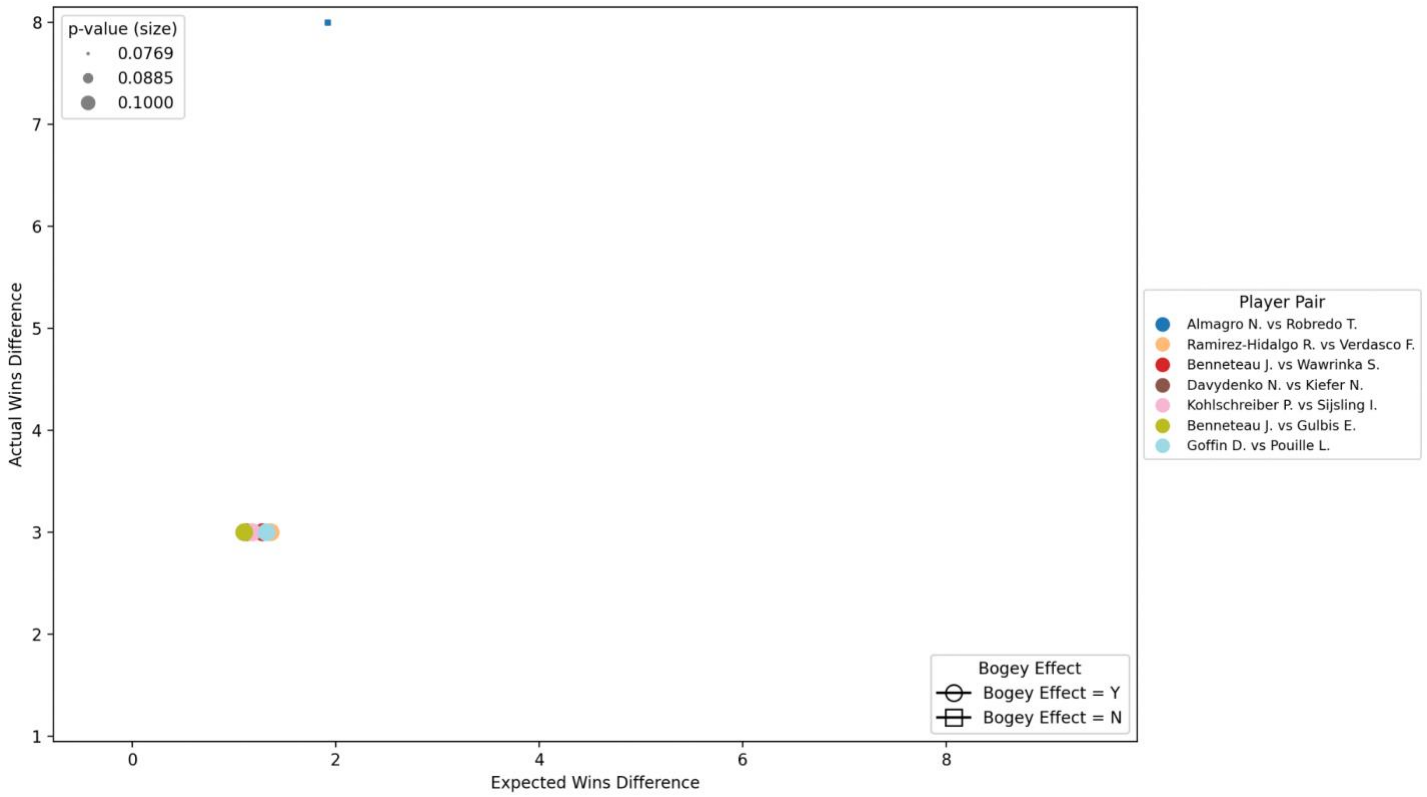
Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Stosur S. vs Zvonareva V.	0.021	2.83/4.17	7/0	Y
Garcia C. vs Ivanovic A.	0.0286	0.96/3.04	4/0	Y
Stephens S. vs Voegele S.	0.0476	4.16/0.84	1/4	N
Safarova L. vs Stosur S.	0.0698	3.67/6.33	8/2	Y
Petkovic A. vs Radwanska A.	0.0769	3.12/4.88	0/8	N
Goerges J. vs Stosur S.	0.0801	1.84/4.16	5/1	Y
Ivanovic A. vs Kuznetsova S.	0.0824	5.1/3.9	9/0	N
Sharapova M. vs Williams S.	0.0867	3.63/6.37	0/10	N
Chakvetadze A. vs Jankovic J.	0.1	0.87/2.13	3/0	Y
Hibino N. vs Stosur S.	0.1	0.59/2.41	3/0	Y
Date Krumm K. vs Kirilenko M.	0.1	0.93/2.07	3/0	N
Ostapenko J. vs Wozniacki C.	0.1	0.67/2.33	3/0	Y
Petkovic A. vs Puig M.	0.1	2.16/0.84	0/3	Y



Supplementary Figure 7: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) bogey player pairs from the ATP non-Grand Slams dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins. The data used to generate this plot is presented below in Supplementary Table 8.

Supplementary Table 8: Statistically significant (with at least 90% confidence) bogey player pairs from the ATP non-Grand Slams dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Garcia-Lopez G. vs Youzhny M.	0.0291	1.9/5.1	6/1	Y
Baghdatis M. vs Garcia-Lopez G.	0.0476	3.28/1.72	0/5	Y
Llodra M. vs Simon G.	0.0476	1.63/3.37	5/0	Y
Almagro N. vs Robredo T.	0.0769	4.47/3.53	8/0	N
Ancic M. vs Blake J.	0.1	0.91/2.09	3/0	Y
Acasuso J. vs Rochus C.	0.1	2.25/0.75	0/3	N
Ramirez-Hidalgo R. vs Verdasco F.	0.1	0.85/2.15	3/0	Y
Benneteau J. vs Wawrinka S.	0.1	0.75/2.25	3/0	Y
Ginepri R. vs Korolev E.	0.1	2.16/0.84	0/3	Y
Almagro N. vs Istomin D.	0.1	2.11/0.89	0/3	Y
Gulbis E. vs Stepanek R.	0.1	0.82/2.18	3/0	Y
Kohlschreiber P. vs Sijsling I.	0.1	2.28/0.72	0/3	Y
Benneteau J. vs Gulbis E.	0.1	0.85/2.15	3/0	N
Querrey S. vs Ramos-Vinolas A.	0.1	2.17/0.83	0/3	Y
Lorenzi P. vs Paire B.	0.1	0.93/2.07	3/0	Y



Supplementary Figure 8: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) bogey player pairs from the ATP non-Grand Slams dataset, with aggregated betting odds-implied probabilities used to compute expected wins. This figure was generated using the data presented below in Supplementary Table 9.

Supplementary Table 9: Statistically significant (with at least 90% confidence) bogey player pairs from the ATP non-Grand Slams dataset, with aggregated betting odds-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Almagro N. vs Robredo T.	0.0769	4.96/3.04	8/0	N
Ramirez-Hidalgo R. vs Verdasco F.	0.1	0.82/2.18	3/0	Y
Benneteau J. vs Wawrinka S.	0.1	0.86/2.14	3/0	Y
Davydenko N. vs Kiefer N.	0.1	2.06/0.94	0/3	Y
Kohlschreiber P. vs Sijsling I.	0.1	2.09/0.91	0/3	Y
Benneteau J. vs Gulbis E.	0.1	0.95/2.05	3/0	Y
Goffin D. vs Pouille L.	0.1	2.16/0.84	0/3	Y

Supplementary Table 10: Expected win distribution violation quantification for the various datasets. For each type of player pair – whether the player pair contains a bogey player or not – average expected wins and average actual wins were calculated and scaled based on the total number of historical matches between the player pair. Taking the difference between these two values provides a means of quantifying the degree to which the expected win distribution is violated. This data was used to generate the plot depicted in Figure 11.

Dataset / method	Player pair type	Average expected win	Average actual wins	Average difference between Expected and Actual wins
ATP / Elo				
	Pair w/o bogey player	0.6952	0.6665	-0.0073
	Pair w/ bogey player	0.7040	0.0172	0.6868
ATP / Odds				
	Pair w/o bogey player	0.6787	0.7034	-0.0247
	Pair w/ bogey player	0.7080	0	0.7080
ATP Grand Slam / Elo				
	Pair w/o bogey player	0.7333	0.7329	0.00044
	Pair w/ bogey player	-	-	-
ATP Non-Grand Slam / Elo				
	Pair w/o bogey player	0.6886	0.6497	0.0389
	Pair w/ bogey player	0.7150	0.01020	0.7048
ATP Grand Slam / Odds				
	Pair w/o bogey player	0.7358	0.7741	-0.0382
	Pair w/ bogey player	0.6904	0	0.6904
ATP Non-Grand Slam / Odds				
	Pair w/o bogey player	0.6676	0.6841	-0.0166
	Pair w/ bogey player	0.7041	0	0.7041
WTA / Elo				
	Pair w/o bogey player	0.6887	0.6592	0.0295
	Pair w/ bogey player	0.7322	0.0282	0.7040
WTA / Odds				
	Pair w/o bogey player	0.6722	0.6925	-0.0203
	Pair w/ bogey player	0.6929	0	0.6929
WTA Grand Slam / Elo				
	Pair w/o bogey player	0.7244	0.7043	0.0200
	Pair w/ bogey player	-	-	-
WTA Non-Grand Slam / Elo				
	Pair w/o bogey player	0.6807	0.6475	0.0331
	Pair w/ bogey player	0.7217	0.0567	0.6650
WTA Grand Slam / Odds				
	Pair w/o bogey player	0.7133	0.7331	-0.0198
	Pair w/ bogey player	-	-	-
WTA Non-Grand Slam / Odds				
	Pair w/o bogey player	0.6626	0.6814	-0.0188
	Pair w/ bogey player	0.6677	0	0.6677

A2.2 PUBLICATION 8: "MULTI-AGENT STATISTICALLY DISCRIMINATIVE SUB-TRAJECTORY MINING AND AN APPLICATION TO NBA BASKETBALL"

> Bunker, R., Duy, V. N. L., Tabei, Y., Takeuchi, I., & Fujii, K. (2024). Multi-agent statistically discriminative sub-trajectory mining and an application to NBA basketball. *Journal of Quantitative Analysis in Sports. 2024 Sep 23.* https://doi.org/10.1515/jqas-2023-0039

This study considered performance at the passage of play level. However, unlike Bunker et al. (2021), which utilised event data, the method proposed in this paper used spatio-temporal tracking data in the form of the trajectories of multiple agents (players and the ball) in National Basketball Association (NBA) Basketball.

Discriminative sub-trajectory mining is a sub-discipline of trajectory mining and data mining, the goal of which is to identify sub-trajectories that are more similar to sub-trajectories in one group and less similar to sub-trajectories in the other group (Le Duy et al., 2020). This study proposed a Multi-Agent Statistically Discriminative Sub-Trajectory Mining (MA-Stat-DSM) method that extended the Stat-DSM method (Le Duy et al., 2020) to the multi-agent setting. The algorithm takes a set of binary-labelled agent trajectory matrices as input. It incorporates Hausdorff Distance to identify sub-matrices statistically significantly discriminating between the two groups of labelled trajectory matrices.

Using the publicly available 2015/16 SportVU NBA tracking data, agent trajectory matrices representing attacks containing the trajectories of five agents (the ball; two offensive players, the shooter and last passer; and two defensive players: the shooter defender and last passer defender), were truncated to correspond to an event-based time interval following the receipt of the ball by the last passer until either a shot was attempted or the ball was turned over. The trajectory matrices were labelled as effective or ineffective based on a definition of attack effectiveness proposed in the current study.

Appropriate parameters for MA-Stat-DSM were determined by iteratively applying the algorithm to all matches involving the two top-placed and bottom-placed teams from the 2015/16 NBA season. The method was then applied to selected matches. It identified and visualised key sub-trajectory matrices representing the parts of plays containing the passes and on- and off-the-ball movements that were most relevant in rendering attacks effective or ineffective.

This study is significant because Multi-Agent Statistically Discriminative Sub-Trajectory Mining (MA-Stat-DSM) was shown to be a potentially useful tool for analysing the spatiotemporal tracking data of multiple agents in team sports and identifying the most relevant portions of a particular play that rendered the play successful or unsuccessful.

This is the accepted manuscript version of the following article, published by De Gruyter in Journal of Quantitative Analysis in Sports on September 23, 2024: Bunker, R., Duy, V. N. L., Tabei, Y., Takeuchi, I., & Fujii, K. (2024). Multi-agent statistical discriminative sub-trajectory mining and an application to NBA basketball. Journal of Quantitative Analysis in Sports. The final published version is available at: https://doi.org/10.1515/jqas-2023-0039

Rory Paul Bunker*, Vo Nguyen Le Duy, Yasuo Tabei, Ichiro Takeuchi, and Keisuke Fujii

Multi-agent statistically discriminative sub-trajectory mining and an application to NBA basketball

https://doi.org/10.1515/jqas-2023-0039

Abstract: Improvements in tracking technology through optical and computer vision systems have enabled a greater understanding of the movement-based behaviour of multiple agents, including in team sports. In this study, a multi-agent statistically discriminative sub-trajectory mining (MA-Stat-DSM) method is proposed that takes a set of binary-labelled agent trajectory matrices as input and incorporates Hausdorff distance to identify sub-matrices that statistically significantly discriminate between the two groups of labelled trajectory matrices. Utilizing 2015/16 SportVU NBA tracking data, agent trajectory matrices representing attacks consisting of the trajectories of five agents (the ball, shooter, last passer, shooter defender, and last passer defender), were truncated to correspond to the time interval following the receipt of the ball by the last passer, and labelled as effective or ineffective based on a definition of attack effectiveness that we devise in the current study. After identifying appropriate parameters for MA-Stat-DSM by iteratively applying it to all matches involving the two top- and two bottom-placed teams from the 2015/16 NBA season, the method was then applied to selected matches and could identify and visualize the portions of plays, e.g., involving passing, on-, and/or off-the-ball movements, which were most relevant in rendering attacks effective or ineffective.

Keywords: team sports, trajectory analysis, tracking data, Hausdorff distance, geographic information systems, spatial information

1 Introduction

The development of tracking technology has increased the availability of trajectory data in various domains, and trajectory mining techniques have been developed and applied in various fields, e.g., in biology to understand animal behavior [24] and to understand pedestrian movements [5]. In team sports, tracking data, obtained from video, wearable devices, or optical systems, has traditionally been used primarily by strength and conditioning staff to analyze athlete movements and demands, e.g., to determine the optimal times to replace players during a match to maximize performance and minimize injuries [6, 21]. However, spatiotemporal tracking data also has potential value for sport performance analysts, who can complement their usual analysis of performance indicators [19], derived from event log data captured in video analysis systems such as SportsCode or Dartfish, with analysis derived from tracking data, for a more holistic understanding of player and team performance. Both the review papers of [33] and [14] highlighted the potential benefits of greater collaboration between sports scientists and computer scientists to explore greater use of spatiotemporal tracking data for performance analysis (in soccer but also team sports in general). In basketball, matches can be decomposed into quarters that are, in turn, decomposed

***Corresponding author: Rory Paul Bunker,** Nagoya University, Email: rorybunker@gmail.com
Vo Nguyen Le Duy, University of Information Technology, Ho Chi Minh City, Vietnam; Vietnam National University, Ho Chi Minh City, Vietnam; and RIKEN Center for Advanced Intelligence Project, Chuo City, Japan. Email: duy.vo@riken.jp
Yasuo Tabei, RIKEN Center for Advanced Intelligence Project. Email: yasuo.tabei@riken.jp
Ichiro Takeuchi, Graduate School of Engineering Mechanical Systems Engineering, Nagoya University. Email: ichiro.takeuchi@mae.nagoya-u.ac.jp
Keisuke Fujii, Graduate School of Informatics, Nagoya University; RIKEN Center for Advanced Intelligence Project; PRESTO, Japan Science and Technology Agency. Email: fujii@i.nagoya-u.ac.jp

into individual plays. There is value to coaches and performance analysts in identifying the most important parts of plays, e.g., the portions of plays that discriminate between effective and ineffective attacks.

Statistically discriminative sub-trajectory Mining (Stat-DSM) [23] is a sub-trajectory mining method (a type of trajectory mining [26, 48] algorithm) that identifies sub-trajectories that statistically significantly discriminate between labeled groups of trajectories of a single agent (hereafter, statistically significantly discriminative is shortened to "SSD" or simply "discriminative"). As well as proposing the Stat-DSM method, [23] demonstrated its applicability on datasets consisting of hurricane and vehicle trajectories. The Stat-DSM method cannot be directly applied in team sports in general because there are multiple trajectories corresponding to the movements of multiple agents (players and the ball). Therefore, in the current study, we propose an extension of Stat-DSM, Multi-Agent (MA) Stat-DSM that aims to identify SSD *sub-matrices*, which consist of the sub-trajectories of multiple agents. To identify SSD sub-trajectories or SSD sub-matrices in the case of Stat-DSM or MA-Stat-DSM, respectively, a distance metric [38] is required, for which we select Hausdorff distance for MA-Stat-DSM (see subsection 4.2 for more detail).

To demonstrate the applicability of the proposed MA-Stat-DSM method, it is applied to the Stats Perform (Chicago, IL, USA) SportVU NBA optical tracking system data from the 2015/16 NBA basketball season, which was preprocessed to contain player and ball trajectories in attacks, which were sub-sampled at a frequency of 5Hz. The trajectories of five agents — two attacking players, two defending players, and the ball — were considered. In particular, each trajectory matrix represents an attack consisting of trajectories of the shooter, the last passer, the shooter defender (closest to the shooter at the time of the shot), and the last passer defender. Each trajectory matrix was labeled based on whether it is an effective or ineffective play, with the effective and ineffective labels computed based on three factors: the position of the shooter on the court (court area/zone), the distance of the shooter from the nearest defender (whether the shot was wide open), and, in the case where a shot is attempted outside the three-point arc, the shooter's historical shot success percentage.

The three main contributions of this study are as follows:

1. A multi-agent statistically discriminative trajectory mining method, MA-Stat-DSM, is proposed that extends Stat-DSM to take the trajectories of multiple agents, in the form of a trajectory matrix, as input and identify the most relevant portion of each attack by obtaining SSD sub-matrices. Unlike machine learning approaches, MA-Stat-DSM does not require complex feature engineering from point coordinates (e.g., by computing, velocities, accelerations, angles, etc.), and its underlying mechanisms are more intuitively understandable compared to black-box deep learning approaches.
2. A novel approach to defining effective and ineffective attacks in basketball is proposed based on the concept of wide-open shots. Each attacking play (trajectory matrix) is labeled as effective or ineffective based on this definition.
3. The proposed method is demonstrated on SportVU NBA trajectory data. In particular, MA-Stat-DSM is applied to the attacks of a specific team in a particular match to identify the portions (sub-matrices) of attacks (agent trajectory matrices) that discriminate between effective and ineffective plays, which could reveal useful post-match insights for coaches and performance analysts.

The remainder of the paper is organized as follows. An overview of related studies is provided in Section 2. Then, in Section 3, we describe the trajectory dataset used in this study, including the proposed computation of effective and ineffective attack labels. Section 4 then describes the proposed MA-Stat-DSM itself. Section 5 provides visualizations of SSD sub-matrices in matches involving, for generality, top- and bottom-performing teams from the 2015/16 season. Finally, Section 6 discusses the obtained results, potential limitations, and avenues for further research.

2 Related Work

Many basketball studies related to tracking data have utilized optical tracking data from the SportVU arena camera system of STATS Perform (prior to 2017 STATS provided tracking data to the NBA) derived from video footage obtained by multiple cameras in the basketball arenas [41]. As mentioned in the introduction, we also use the SportVU data, which is described further in subsection 3.1.

Statistical methods, e.g., cluster analysis and analysis of variance (ANOVA), have been applied to tracking and non-tracking data to construct performance indicators and profiles related to scoring, passing, defensive and all-around game roles and behavior. Network-based models [37] have also been applied to tracking data to enable the enhanced evaluation of individual player skills/performance and prediction of team performance in basketball that can surpass traditional statistics-based approaches. Deep learning-based computer vision techniques have also been proposed to analyze tracking data in basketball, e.g., to classify player and ball movements from video and to analyze passing relationships [45]. Tracking data in the 3-s lead-up to three-point attempts has been used to analyze movement patterns that create "open shots" (where the nearest defender to the shooter is at a distance of at least 6 feet away) and how these can impact performance [25]. As mentioned, the concept of open shots comprises part of the effective/ineffective attack label definition that is proposed in the current study. Tracking data in the lead-up to three-point shots has been converted into sequences, and recurrent neural networks, a deep learning model, have been used to predict three-point shot success/failure [35]. A long short-term memory (LSTM) network [17] with neural embedding and deep feature representation was proposed by [36], who formulated a multi-class sequence classification problem that uses spatiotemporal tracking data as input. The approach estimates the probabilities of actions taken by players at the end of possessions, which can determine expected points at each point in time during an attacking play and can, in turn, be used to evaluate so-called micro-actions in terms of their contribution to the success of a possession.

Strategy identification and classification in basketball is another area of study that uses tracking data. For instance, neural and recurrent neural networks, which are able to handle sequential data that are of varying lengths, have been applied to SportVU trajectories that were converted into image representations for classification of attacking plays and sequence prediction, and to investigate whether the model could classify offensive plays in a subsequent season [43]. One of the primary contemporary offensive strategies used in the NBA is the pick-and-roll/ball screen. A machine learning classification model on top of a rule-based algorithm was used by [28] to identify on-ball screens from SportVU tracking data from 21 quarters across 14 matches in the 2012/13 NBA season. Building on this work, [27] applied a supervised machine learning classifier to SportVU data from the 2011/12 to 2014/15 seasons to automatically recognize defensive strategies employed against ball screens. Machine learning techniques, e.g., k-nearest neighbors, decision trees, and support vector machines, have been applied to various tracking data-derived features such as player velocities, inter-player distances, player movement vector similarity, and defensive zones, to classify defensive (switch and trap) strategies used against pick-and-rolls [42]. Using player tracking system data from 1,230 regular season matches in the 2013-2014 season, [34] used discriminant analysis to distinguish between the performance of all-star and non all-star players in NBA basketball, and identified role-based performance profiles of players using k-means clustering. Active learning with neural networks is another approach that has been proposed to circumvent time-consuming annotation by domain experts to identify the pick-and-roll offensive strategy using tracking data [2]. Semi-supervised learning has also recently been proposed for the classification of ball screen plays from SportVU tracking data [50].

Methods that track player and ball movements can be useful for performance evaluation and strategy identification/classification. However, it is not possible to holistically analyze team performance without considering the movements and interactions of all players as a group [15]. Multi-agent methods that consider the movements of agents including the players and ball are, therefore, important in this context, and deep learning approaches such as bidirectional LSTM and mixture density networks have been used for trajectory prediction [47] and assisting in decision-making regarding the optimal locations and times to make a shot. Tensors [31] and transformers [3] are other deep learning approaches that have been used to

model multi-agent spatiotemporal data in basketball, and graph-based representations [32] have also been proposed in other sports (soccer).

Of relevance to the current study are methods that identify relevant parts of plays, e.g., those that discriminate between good and bad outcomes or different types of periods of play. For instance, [8] developed an algorithm that discriminates between active and inactive periods of play using trajectory data derived from sensor tracking data. Video clip data has been converted into player trajectory/action representations to analyze offensive strategies of differing duration in basketball, and dynamic time warping used to compute the similarity of video clips, and a largely unsupervised approach with clustering used to divide training data and Gaussian mixture regression employed to robustly model discriminative between-label variations [7]. A multi-agent neural network-based approach based on an attention mechanism using features related to multi-agent movements e.g., the distances between agents and objects, was recently proposed to identify trajectory segments that are correlated with effective/ineffective and scoring/non-scoring attacks [49]. The method proposed in the current study has two main advantages over [49]: first, it does not require the extraction of movement-related features from the original trajectory data, and second, it is more intuitive compared to the black-box nature of deep learning methods. Discriminative methods have also been used in other sports to identify discriminative patterns from event sequence data. For example, discriminative sequential pattern mining has been applied to event sequences derived from event log data in rugby to identify subsequences (patterns) that discriminate between scoring and non-scoring plays [4]. "Interestingness" measures from mined frequent sequential patterns have also been obtained during training in cycling [18].

3 Data

In this section, the preprocessed SportVU NBA trajectory dataset, to which the proposed method will be applied, is described in subsection 3.1. Then, in subsection 3.2, the approach for the computation of the effective and ineffective attack labels is described.

3.1 Trajectory data

In this study, we used attack sequences from 600 regular season games from the 2015/2016 NBA season, which was originally sourced from GitHub from the 2015-2016 NBA raw SportVU game logs (https://github.com/neilmj/BasketballData/tree/master/2016.NBA.Raw.SportVU.Game.Logs). The dataset originally contained the trajectories of 11 agents, five players on each opposing team and the ball. We considered five of these agents: the ball and two players from each opposing team (the shooter, the shooter defender, the last passer, and the defender of the last passer).

Since scoring prediction is generally difficult, and non-linear data-driven approaches may sometimes not be interpretable (e.g., [10, 11]), we defined the label to be based on whether or not a particular play was an "effective attack" rather than whether or not points were scored in that play. The definition/computation of effective and ineffective attacks will be provided in detail in subsection 3.2.

There were a total of 45,307 attacks, sub-sampled at 5 Hz; therefore, the time between consecutive points is always the same (0.2s). As a result, although the raw data is spatiotemporal, only its spatial dimension needs to be considered. In our dataset, there were 18,021 shot successes, 20,155 shot failures, 22,159 effective attacks, and 23,148 ineffective attacks. This dataset was already split into attacks, but as a preprocessing step, we removed the duplicate attacks and trimmed the start and end times. The probabilities of scoring, given the attack was effective and ineffective, were 0.466 and 0.333, respectively. A Chi-squared test with a 2-by-2 contingency table with scoring as the columns and effectiveness as the rows yielded a p-value very close to 0, indicating that the difference between these two values (0.466 and 0.333) was statistically significant.

3.2 Effective and ineffective attack labeling

In this subsection, we describe our approach to computing effective and ineffective attacks, which are used as the trajectory matrix (attack) labels. Due to the differing shooting abilities of individual players and other stochastic factors, evaluating team movements based on scores alone may not provide a holistic view of a "good" attacking play. Indeed, it could be argued that the tactics and strategy of a coach and team are most influential up until the point at which there is a good scoring opportunity, i.e., the creation of a chance to attempt a shot. It is then the skills and form of the individual player that determines whether this scoring opportunity is actually converted into points. We consider a good scoring opportunity in basketball to be a shot that is attempted in a context in which there is a high expected probability of scoring, based on historical attempted and successful shots. Therefore, we compute an interpretable and simple indicator from available statistics, based on frequencies, to evaluate whether a player makes an effective shot attempt, rather than based on a label with only successful shots or learning-based score prediction.

From the available statistics, we focused on two basic factors for effective attacks at an individual player level: shot zone on the court, and the distance between a shooter and the nearest defender. These two factors are considered to be important for basketball shot prediction [10, 11, 13]. The spacing among players, derived from tracking data, has been shown to have an impact on team performance [29]. In the NBA advanced stats (https://www.nba.com/stats/players/shots-closest-defender/), we have access to the probabilities of successful shots attempted in each zone, as well as distances, for each player. The shot zones are partitioned into four areas: restricted, in-the-paint, mid-range, and 3-point areas. The restricted area is defined as the area within a radius of 2.44 m (the distance between the side of the rectangle and the hoop) from the hoop. The in-the-paint area is defined as the area within a radius of 5.46 m (the distance between the hoop and the farthest vertex of the rectangle) from the hoop. The three-point area is defined as the area that is outside of the 3-point line. The mid-range is the remaining area. The distance of the shooter from the nearest defender is categorized into four ranges: $0-2$ feet, $2-4$ feet, $4-6$ feet, and 6+ feet.

We define an effective attack as one that meets the following criteria:

- The position of the shooter in the restricted area is effective at any distance because of their proximity to the hoop (despite a defender often being located near the shooter).
- The position of the shooter in-the-paint and mid-range areas is effective at a distance of six feet or more from the nearest defender (this range is regarded as "open" in the NBA advanced stats).
- The position of the shooter in the 3-point area is effective when a player with a shot success probability of at least 0.35 attempts a shot at a distance of 6 feet or more from the nearest defender (because some players do not shoot tactically). A limitation of the dataset used is that the shot success probability data needed to be aggregated over all three-point shots attempted, not only those at which the defender is more than 6 feet away. Thus, it should be noted that a player's actual undefended shot success probability is probably higher than the recorded statistic.

The probability of 0.35 is determined by the simple idea of a 3-point shot being "not bad." If we assume that 50% of 2-point shots are successful (this is determined subjectively, but is not unrealistic), we can select a 3-point shot if more than 33% of 3-point shots are successful. Therefore, we determined the threshold as 35%. Of course, this is a rough estimation and ideally, this should be customized for each team's strategy, but this is beyond the scope of the current study.

Based on the statistics in the 2014/2015 season and the tracking data, we computed the probabilities of successful shots for each zone and the distances for each player. We computed the probabilities of players who had attempted less than 10 shots based on those of players of the same position (i.e., guard, forward, center, guard/forward, and forward/center, based on the registration in the NBA 2014/2015 season). Note that the 10-shot threshold is not based on total shot attempts, but applies to 2- and 3-point shot attempts for each range of distances between the shooter and shooter defender (0m-2m,2-4m,4-6m,6m+). Using a higher threshold of 30 was found to greatly reduce the number of attacks available for the analysis. It should be noted that certain characteristics of a good shot can differ depending on the court location and context, e.g., for 2- and 3-pointers. Note that, unfortunately, we could access those for only two areas (the 2- and 3-point areas) with four distance categories. Thus, we computed the shot success probabilities in the

restricted, in-the-paint, and mid-range areas using those in the 2-point area. To adjust the 3-point shot probabilities for shots attempted a long distance from the 3-point line, we linearly reduced the probabilities by 0.2 at 12.73 m (the distance between the hoop and the half-court line).

In a naïve approach, the results obtained using trajectories consisting of the entire attack segments would not be interpretable and would not provide useful results because the roles of the players are not aligned with the order of the players in the trajectory data. To extract meaningful information, we focus on the trajectories of five agents (the ball and four players) in the interval from when the last passer receives the ball until the shooter makes a shot (interval t2 to t0 shown in Figure 1). When a shot is not attempted in a particular play, the end time of the trajectory is determined as the time at which a turnover occurs (i.e., when the defensive team comes to be in possession of the ball).

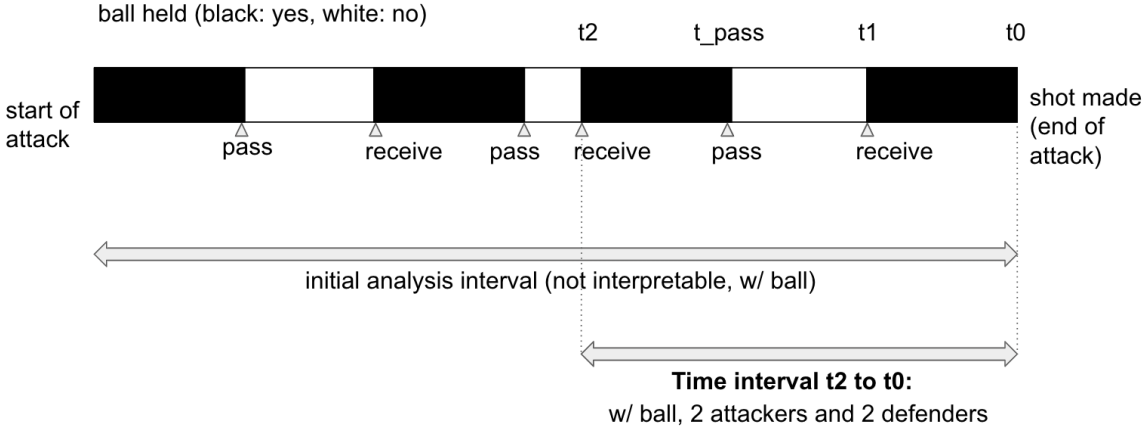


Fig. 1: The attacks were cropped to consider the time interval from t2 to t0, i.e., the trajectories of the ball, 2 attackers (shooter and last passer), and 2 defenders (shooter defender, last passer defender) during the time from which the last passer receives the ball until a shot is made or the ball is turned over (t1 is the time at which the shooter receives the ball).

4 Method

In this section, we first describe in subsection 4.1 the benefits of applying MA-Stat-DSM to trajectory matrices (attacks) over Stat-DSM to individual agent trajectories, and how we resolve the resulting role assignment problem. Then, we outline our problem setup and provide relevant definitions in subsections 4.2 and 4.3, respectively. Finally, we provide the pseudocode for the MA-Stat-DSM algorithm.

4.1 Role Assignment

One naïve approach would be to apply Stat-DSM to the trajectories of each agent. However, this approach would result in losing potentially important spatiotemporal information about multi-agent interactions and would also require more computational time. Furthermore, there may be interactions between players that would not be captured if the Stat-DSM method was applied to each individual agent trajectory and the results were combined. Therefore, we apply MA-Stat-DSM to the set of multi-agent trajectories, which requires a less-than-proportional increase in computational time with respect to the number of agents considered.

Consequently, a role-assignment problem occurs in multi-agent data processing. Since the players are usually ordered randomly in the raw data, meaningful roles such as position (e.g., guard, forward, and center) are ignored. Generally, this problem can be solved by a linear assignment problem [30], e.g., using a Gaussian hidden Markov model [12, 22]. However, since this approach is data-driven, it lacks the interpretability of each assigned role. To retain the interpretability of results, which is a concern along with computational cost, we assign meaningful roles (e.g., shooter, shooter defender, etc.) in a rule-based manner. We selected the five agents (ball, shooter, shooter defender, last passer, and last passer defender) because considering all trajectories may be too diverse for the model to extract useful information. In general, this is a multi-agent role assignment problem for an unsorted, diverse dataset, which can be avoided by using only predetermined roles about these four players and the ball. As mentioned, a data-driven approach was also considered, but we prioritized interpretability. It is more difficult to determine the roles in a fixed manner as the number of players increases (e.g., it is difficult to generally determine the roles other than those of the shooter and passer), and fewer players may be less informative in this analysis. Therefore, we consider that all trajectories with all player role assignments will lose some information in turn; thus, we selected these five agents.

4.2 Problem Setup

The key differences between the proposed MA-Stat-DSM method and the original Stat-DSM method are that the labelled item in Stat-DSM is a trajectory whereas it is a trajectory matrix in MA-Stat-DSM, and instead of aiming to identify SSD sub-trajectories as in Stat-DSM, in MA-Stat-DSM, we aim to identify SSD sub-matrices (which are comprised of agent sub-trajectories). Euclidean distance cannot be used to compute the distance between matrices that consist of multi-agent trajectories, so we incorporate an efficient implementation of Hausdorff distance [20], which was proposed by [39], and is available on GitHub (https://github.com/mavillan/py-hausdorff). While Euclidean distance is used in Stat-DSM to compute the distance between two (sub)-trajectories of a single agent, in the multi-agent setting, the distance between matrices of agent trajectories, which are of differing lengths across attacks, needs to be determined. Inspired by its use in the point set distance metric approach in multiple-instance learning [16], we incorporate Hausdorff distance in MA-Stat-DSM to determine the distance between (sub)-matrices of differing lengths.

There is assumed to be a set of K agents, each of which has a trajectory in all N matrices. Each trajectory matrix has K rows and a length corresponding to the number of coordinates in each agent trajectory. The lengths of each of the agent trajectories are the same within a trajectory matrix. The length of the agent trajectories represents the number of columns in the trajectory matrix. Therefore, the lengths of each of the agent trajectories being the same within a trajectory matrix implies that there are no point coordinates that have values. On the other hand, different trajectory matrices will generally be of different lengths, i.e., have different numbers of columns.

In the context of the current study related to basketball, the agents represent the ball or a player, and the trajectory matrices represent plays. In particular, there are $|K| = 5$ agents: $K = \{$ball, shooter, shooter defender, last passer, last passer defender$\}$. Each of the K agents has a trajectory that is of the same length in each trajectory matrix (play). A single agent in the set of all agents is denoted by $k \in K$. The ith play/trajectory matrix, of which there are N in total, consists of the trajectories of each agent in that play and can be represented as a $|K| \times m_i$ matrix, where m_i is the length of play i. In general, the lengths of each play differ, i.e., $m_i \neq m_j$ for $i \neq j$.

In this study, we apply the MA-Stat-DSM method to the attacks of a particular team in a specific match. The method was initially applied to the attacks of a particular team across the whole 2015/16 season, but the computational complexity was found to be prohibitive. Therefore, we considered instead the attacks of teams in a particular match and then iterated over the matches in the season. Therefore, we aimed to use MA-Stat-DSM to identify SSD sub-matrices, i.e., portions of attacks, which discriminated between team T's effective and ineffective attacks in match M.

4.3 Definitions

In this subsection, we provide some definitions and notation for Hausdorff distance, trajectory matrices, sub-matrices, distance, ε-neighborhood, and support, partly based on the definitions provided by [46] and [23].

Hausdorff distance. The Hausdorff distance between two sets of instances (trajectory matrices in our case) is the aggregation of the base distances between the instances in each matrix. Euclidean distance, Manhattan, or Chebyschev distance are commonly used as base distances; we used Euclidean distance in this study. Two matrices, X and Y, are within a Hausdorff distance of $dist_H$ if and only if every point in X is within distance $dist_H$ of at least one point in Y, and every point in Y is within distance $dist_H$ of at least one point in X. In particular, the Hausdorff distance, $dist_H(X, Y)$, between two point sets (matrices) is, in general:

$$dist_H(X, Y) = \max\{h(X, Y), h(Y, X)\}, \tag{1}$$

where $h(X, Y) = \max_{x \in X} \min_{y \in Y} dist(x, y)$.

Trajectory matrix. The trajectory of agent k in matrix i is a finite sequence of m_i points: $T_{i,k} = \{(x_{1,k}, y_{1,k}), (x_{2,k}, y_{2,k}), ..., (x_{m_i,k}, y_{m_i,k})\}$. Trajectory matrix i contains the trajectories of all K agents in a specific play, has m_i columns, and is denoted $\boldsymbol{T}_{i,K}$. There are N trajectory matrices in a specific match M, each of which takes a label from $g_i = \{+1, -1\}$, and $G_+ = \{\boldsymbol{T}_{i,K} \mid g_i = +1\}$ and $G_- = \{\boldsymbol{T}_{i,K} \mid g_i = -1\}$ denote the groups of trajectory matrices with labels +1 and -1, respectively. In the current study, as mentioned, a trajectory matrix represents an attack, which is labeled as either effective or ineffective.

Trajectory sub-matrix. A sub-matrix is denoted $\boldsymbol{T}_{i,K}^{(s,e)}$, and is a sequence of consecutive columns within the trajectory matrix $\boldsymbol{T}_{i,k}$, starting from column index s and ending at e, with a fixed number of $|K|$ rows. The length of sub-matrix i is $|\boldsymbol{T}_{i,K}^{(s,e)}| \geq L$, where L, the minimum length (number of columns) of the sub-matrix, is a user-selected parameter of MA-Stat-DSM. The notation $\boldsymbol{T}_{i,K}^{(s,e)} \sqsubseteq \boldsymbol{T}_{i,K}$ indicates that $\boldsymbol{T}_{i,K}^{(s,e)}$ is a sub-matrix of $\boldsymbol{T}_{i,K}$.

Distance metric between sub-matrices. Using the general definition of Hausdorff distance above (Equation 1), the Hausdorff distance, $dist_H(\boldsymbol{T}_{i,K}^{(s,e)}, \boldsymbol{T}_{i',K}^{(s',e')})$, between two agent trajectory sub-matrices, $\boldsymbol{T}_{i,K}^{(s,e)} = \{T_{i,1}^{(s,e)}, T_{i,2}^{(s,e)}, ..., T_{i,|K|}^{(s,e)}\}$ and $\boldsymbol{T}_{i',K}^{(s',e')} = \{T_{i',1}^{(s,e)}, T_{i',2}^{(s,e)}, ..., T_{i',|K|}^{(s,e)}\}$, is:

$$dist_H(\boldsymbol{T}_{i,K}^{(s,e)}, \boldsymbol{T}_{i',K}^{(s',e')}) = \max h(\boldsymbol{T}_{i,K}^{(s,e)}, \boldsymbol{T}_{i',K}^{(s',e')}), h(\boldsymbol{T}_{i',K}^{(s',e')}, \boldsymbol{T}_{i,K}^{(s,e)}), \text{ where}$$

$$h(\boldsymbol{T}_{i,K}^{(s,e)}, \boldsymbol{T}_{i',K}^{(s',e')}) = \max_{T_{i,k}^{(s,e)} \in \boldsymbol{T}_{i,K}^{(s,e)}} \min_{T_{i',k}^{(s',e')} \in \boldsymbol{T}_{i',K}^{(s',e')}} dist(T_{i,k}^{(s,e)}, T_{i',k}^{(s',e')})$$

and

$$h(\boldsymbol{T}_{i',K}^{(s',e')}, \boldsymbol{T}_{i,K}^{(s,e)}) = \max_{T_{i',k}^{(s',e')} \in \boldsymbol{T}_{i',K}^{(s',e')}} \min_{T_{i,k}^{(s,e)} \in \boldsymbol{T}_{i,K}^{(s,e)}} dist(T_{i',k}^{(s',e')}, T_{i,k}^{(s,e)})$$

Trajectory sub-matrix ε-similar-neighborhood and support. The ε-similar-neighborhood for each sub-matrix is the set of sub-matrices within a Hausdorff distance of ε, and is given by:

$$N_\varepsilon(\boldsymbol{T}_{i,K}^{(s,e)}) := \{\boldsymbol{T}_{i',K}^{(s',e')} \mid \text{dist}_\text{H}(\boldsymbol{T}_{i,K}^{(s,e)}, \boldsymbol{T}_{i',K}^{(s',e')}) \leq \varepsilon\},$$

where ε is the distance threshold, a user-selectable parameter of MA-Stat-DSM.

The support of $\boldsymbol{T}_{i,K}^{(s,e)}$ with respect to a subset of sub-matrices $G_m \subseteq [n]$ (where $[n]$ denotes the set of all trajectory matrices in the dataset) is:

$$\sup_{G_m}(\boldsymbol{T}_{i,K}^{(s,e)}) := |\{i' \in G_m \mid \exists\, \boldsymbol{T}_{i',K}^{(s',e')} \sqsubseteq \boldsymbol{T}_{i',K}, \boldsymbol{T}_{i',K}^{(s',e')} \in N_\varepsilon(\boldsymbol{T}_{i,K}^{(s,e)})\}|$$

This support represents the number of sub-matrices in G_m that contain at least one sub-matrix with distance from sub-matrix $\boldsymbol{T}_{i,K}^{(s,e)} \leq \varepsilon$.

As mentioned, in this study, we consider the set of all trajectory matrices (attacks) of a particular team in a specific match, so when referring to "all trajectory matrices in the dataset," in our problem setup this means "all attacks by team T in match M."

MA-Stat-DSM Algorithm. The flow of MA-Stat-DSM (Algorithm 1) pseudocode is relatively similar to Stat-DSM, so for full detail, we refer the reader to [23]. The main changes are that Euclidean distance is replaced by Hausdorff distance ($dist_H$ on lines 27 and 28 of the pseudo-code), trajectory and sub-trajectory are replaced with trajectory matrix and sub-matrix, respectively, and a fixed set of K agents is considered. Figure 2 depicts the main steps of the MA-Stat-DSM algorithm and the lines corresponding to Algorithm 1.

The general steps of the algorithm, as shown in Algorithm 1 and Figure 2, can be described as follows. The initialization occurs in lines 2 to 5. Here, the labels are computed for all permuted datasets ($B = 1000$ times, and the minimum p-values of each bth permuted dataset, $p_{min}^{(b)}$ are set to be α. Lines 6 to 11 involve the simultaneous extraction of sub-matrices and estimation of the p-value null distribution. Specifically, sub-matrices are extracted as nodes of a tree, with child nodes representing successively longer sub-matrices than the parent node and preceding child nodes (the first node of each tree branch is a sub-matrix with length L). At line 8, the ε-similar neighborhood is determined. Then, at line 9, the ProcessNext() function is called and re-called iteratively.

The ProcessNext() function processes one sub-matrix each time it is called, and simultaneously performs sub-matrix extraction and updates the null distribution. Within the ProcessNext() function, line 16 involves sorting the list of current minimum p-values in ascending order. Line 17 computes the lower bound of the current sub-matrix/tree node. If the pruning criterion at line 18 is satisfied, the current branch is no longer explored (lines 18 – 20), i.e., processing stops. Otherwise, processing continues, and the minimum p-value distribution is updated (lines 21 – 25). The child nodes (longer trajectory sub-matrices) are continued to be explored and the null distribution continues to be updated, and the ProcessNext() function is re-called on longer sub-matrices (lines 26 – 32) until the $\alpha B + 1$ smallest minimum p-values have been computed.

Finally, in lines 12 and 13, the optimal adjusted significance threshold, δ^*, is calculated, and the obtained SSD sub-matrices, which have p-values $< \delta^*$, are output on line 14.

The Python code for MA-Stat-DSM is available on GitHub (see the Appendix for the URL).

4.4 Basketball-specific Example

Although specific plays will be interpreted for other matches in the results section, here, we describe an example of the problem setup and how MA-Stat-DSM is applied to the set of team attacks from a specific match, as depicted in Figure 3.

The attacks of the Los Angeles Lakers from their 15 November 2015 match against the Detroit Pistons are shown, with the effective attacks (G_+) and ineffective attacks (G_-) shown in the top and bottom panels, respectively. The $N = 12$ plays, which are represented as $|K|$-by-m_i trajectory matrices, are shown based on the time interval from when the last passer receives the ball to when a shot is attempted, t2 to t0 (Figure 1). There are five effective attacks, i.e., $|G_+| = 5$, and seven ineffective attacks, i.e., $|G_-| = 7$, based on our proposed definition of effectiveness. That is, each of the top five matrices is labelled effective (+1) and each of the bottom seven matrices has an ineffective (-1) label. Each of the twelve attacks represents an agent trajectory matrix, $\boldsymbol{T}_{i,K}$, which consists of the contemporaneous trajectories ($T_{i,k}$) of each of the $|K| = 5$ agents considered, i.e., $K = \{$ball, (Lakers) shooter, (Lakers) last passer, (Pistons) shooter defender, (Pistons) last passer defender$\}$. The number of rows in the i-th agent trajectory matrix corresponds to the number of agents, $|K| = 5$, and the number of columns in each agent trajectory matrix corresponds to the length of the trajectories, m_i (number of point coordinates). As mentioned, all agent trajectories within the same trajectory matrix have the same length ($m_{i,k}$ is the same for all $k \in K$, but the length of trajectories across attacks differs. Specifically, due to the high frequency of the SportsVU data, it is unlikely (but not impossible) that the length (number of columns) of one play is the same as another (i.e., in general, $m_i \neq m_{i'}$ for $i \neq i'$).

In this problem setup, MA-Stat-DSM (described in the next section) can be applied to this set of twelve labelled agent trajectory matrices to identify whether there is a portion of the ith attack, in the form of an SSD sub-matrix, $\boldsymbol{T}_{i,K}^{(s,e)}$, which discriminates between the effective and ineffective labels. The discriminative sub-matrices again have $|K| = 5$ rows, and contain the SSD contemporaneous sub-trajectories, $T_{i,K}^{(s,e)}$, of

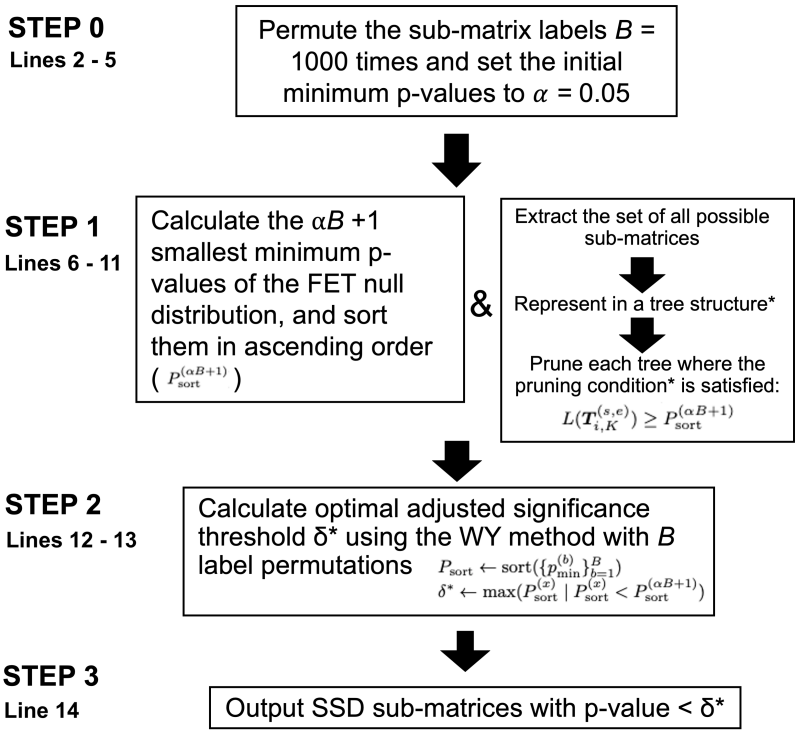


Fig. 2: Main steps of the MA-Stat-DSM algorithm with the corresponding lines in Algorithm 1. Although we provide a brief description of the statistical testing and pruning process of Stat-DSM and MA-Stat-DSM in the appendix to this paper, please refer to sections 3 and 4 in [23] for a full description of the flow of Stat-DSM, which also is applicable to MA-Stat-DSM but with the concepts of trajectory and sub-trajectory replaced with trajectory matrix and sub-matrix, respectively, and Euclidean distance replaced by Hausdorff distance.

Algorithm 1: Multi-agent statistically discriminative sub-trajectory mining (MA-Stat-DSM)

Input: Set of K agents (fixed), Trajectory matrix dataset $\boldsymbol{D} = \{\boldsymbol{T}, \boldsymbol{g}\}$, distance threshold ε, minimum length L, number of permutations B, and significance level α.

Output: Statistically discriminative sub-matrices

procedure *Main* ()

// Initialization

for $b \leftarrow 1$ to B **do**

$\boldsymbol{D}^{(b)} \leftarrow \{\boldsymbol{T}, \text{permute}(\mathbf{g})\}$

$p_{\min}^{(b)} \leftarrow \alpha$

end

// Extract sub-matrices and estimate the null distribution

for each $\boldsymbol{T}_{i,K} \in \boldsymbol{T}$ **do**

for each length-L sub-matrix $\boldsymbol{T}_{i,K}^{(s,e)} \sqsubseteq \boldsymbol{T}_{i,K}$ **do**

Compute $N_\varepsilon(\boldsymbol{T}_{i,K}^{(s,e)})$

ProcessNext $(\boldsymbol{T}_{i,K}^{(s,e)},\ N_\varepsilon(\boldsymbol{T}_{i,K}^{(s,e)}))$

end

end

// Calculate the adjusted significance level δ^*

$P_{\text{sort}} \leftarrow \text{sort}(\{p_{\min}^{(b)}\}_{b=1}^{B})$

$\delta^* \leftarrow \max(P_{\text{sort}}^{(x)} \mid P_{\text{sort}}^{(x)} < P_{\text{sort}}^{(\alpha B+1)})$

// Statistically discriminative sub-matrices

Output the sub-matrices with p-values $< \delta^*$

function *ProcessNext* $(\boldsymbol{T}_{i,K}^{(s,e)},\ N_\varepsilon(\boldsymbol{T}_{i,K}^{(s,e)}))$

$P_{\text{sort}} \leftarrow \text{sort}(\{p_{\min}^{(b)}\}_{b=1}^{B})$

Compute $L(\boldsymbol{T}_{i,K}^{(s,e)})$

if $L(\boldsymbol{T}_{i,K}^{(s,e)}) \geq P_{\text{sort}}^{(\alpha B+1)}$ **then**

return

end

for $b \leftarrow 1$ to B **do**

if $L(\boldsymbol{T}_{i,K}^{(s,e)}) < p_{\min}^{(b)}$ **then**

$p_{\min}^{(b)} \leftarrow \min\{\ p_{\min}^{(b)}, p^{(b)}(\boldsymbol{T}_{i,K}^{(s,e)})\ \}$

end

end

for each $\boldsymbol{T}_{i',K}^{(s',e')} \in N_\varepsilon(\boldsymbol{T}_{i,K}^{(s,e)})$ **do**

$d \leftarrow \text{dist}_{\text{H}}(\boldsymbol{T}_{i',K}^{(s',e'+1)}, \boldsymbol{T}_{i,K}^{(s,e+1)})$

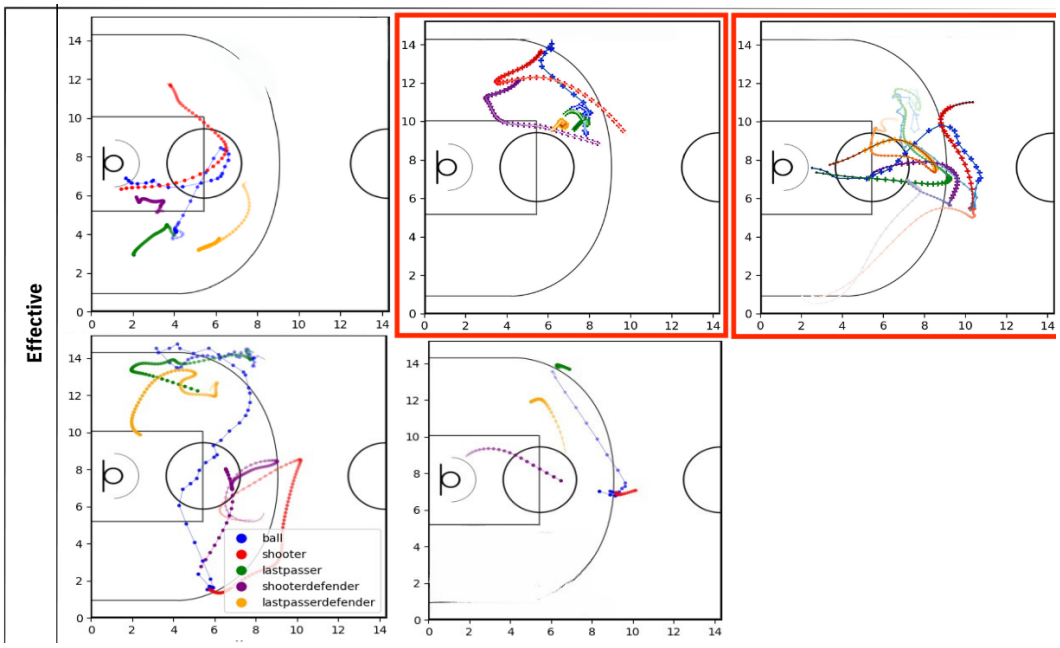
if $d \leq \varepsilon$ **then**

Add $\boldsymbol{T}_{i',K}^{(s',e'+1)}$ into $N_\varepsilon(\boldsymbol{T}_{i,K}^{(s,e+1)})$

end

end

ProcessNext$(\boldsymbol{T}_{i,K}^{(s,e+1)},\ N_\varepsilon(\boldsymbol{T}_{i,K}^{(s,e+1)}))$



(a) Effective attacks

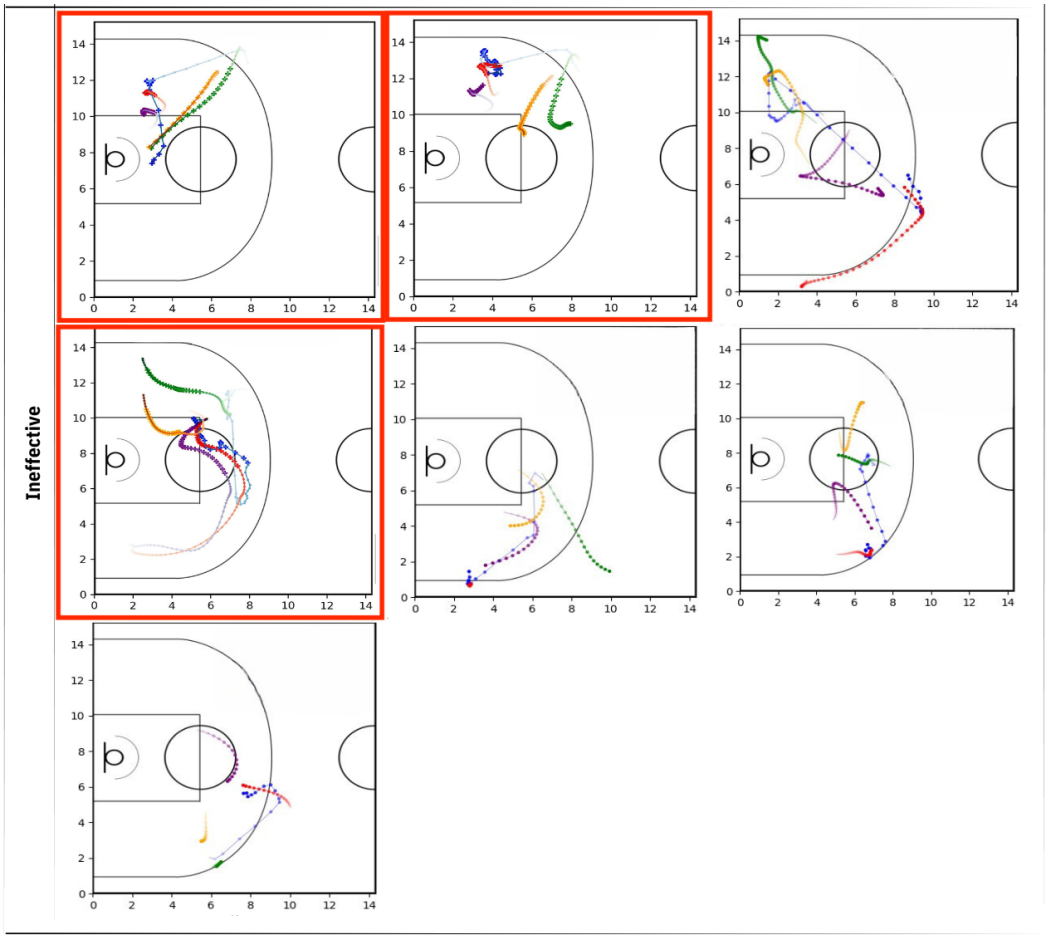
Fig. 3: Lakers attacks from their 15 November 2015 match against the Detroit Pistons. Effective attacks (a) and ineffective attacks (b) are defined based on our proposed definition of effectiveness. Time progression is displayed by the trajectories' colour transitioning from light to dark. The discriminative portions of attacks, i.e., the discriminative (SSD) sub-matrices, are denoted by plus signs. Attacks containing SSD sub-matrices are enclosed in a red rectangle.

each agent $k \in K$. The SSD sub-matrices are denoted by plus signs. As can be observed in Figure 3a and Figure 3b, when applying MA-Stat-DSM (with a distance threshold $\varepsilon = 4$ and a minimum length of $L = 4$) to the 12 Laker attacks from their 15 November 2015 match against the Pistons, five attacks (agent trajectory matrices) contained SSD sub-matrices. In particular, two of the effective plays had discriminative sub-matrices (Figure 3a) and three of the ineffective plays contained discriminative sub-matrices (Figure 3b).

SSD sub-matrices in an effective agent trajectory matrix indicate the portion of the play that was relevant in rendering the attack effective rather than ineffective. Similarly, SSD sub-matrices in an ineffective agent trajectory matrix indicate the portion of the play that was relevant in rendering the attack ineffective rather than effective. Note that in some cases, e.g., the second effective play from the right, the discriminative portion of the play comprised all of (or nearly all of) the agent trajectory matrix, i.e., the SSD sub-matrix and agent trajectory matrix are the same, indicating that the whole passage of play in this time interval was relevant in rendering this attack effective rather than ineffective.

5 Results

In this section, we first outline the experimental setup in terms of the parameters selected for MA-Stat-DSM in subsection 5.1. Then, we provide examples of effective and ineffective discriminative SSD sub-matrix results that were obtained for matches involving two of the top teams from the 2015/2016 NBA competition (Cleveland and Golden State) and the bottom teams from the Eastern and Western Conferences (the Philadelphia 76ers and Los Angeles Lakers) and interpret the results from a practical perspective.



(b) Ineffective attacks

Fig. 3: Lakers attacks from their 15 November 2015 match against the Detroit Pistons. Effective attacks (a) and ineffective attacks (b) are defined based on our proposed definition of effectiveness. Time progression is displayed by the trajectories' colour transitioning from light to dark. The discriminative portions of attacks, i.e., the discriminative (SSD) sub-matrices, are denoted by plus signs. Attacks containing SSD sub-matrices are enclosed in a red rectangle.

5.1 Experimental Setup

Although there was limited *a priori* knowledge as to what the appropriate parameter setting for MA-Stat-DSM should be for the basketball dataset we are using, the paper in which Stat-DSM was proposed [23], which applied the proposed method to datasets consisting of vehicle and hurricane trajectories, was used as a starting point. While these datasets involved greater distances than those on a basketball court, to balance this, the data was also much lower frequency than the NBA trajectory data. Thus, as a starting point, we consider the range of parameters specified for Stat-DSM by [23]. Ultimately, [23] selected a significance level, α, of 0.05 and the number of permutations, B, to be 1,000, and we also selected these values in the present study.

The MA-Stat-DSM parameter values we used are shown in Table 1. MA-Stat-DSM was applied to the trajectory data, with a statistical significance level of 0.05, $B = 1000$, distance thresholds of 1.5 and 4, and minimum lengths of 5, 8 and 10. The data preprocessing parameters used were $|K| = 5$ agents and a time interval from t2 to t0 (Figure 1).

Table 2 shows the number of SSD matrices within attacks (and the number of distinct matches containing those attacks) obtained with the parameter settings in Table 1. It can be observed that, with a distance threshold of 1.5, only very few SSD sub-matrices were obtained by MA-Stat-DSM compared to when a distance threshold of 4.0 was used. A distance threshold of 1.5 means that a play/attack, i.e., trajectory matrix that consists of the five agent trajectories in our case, is within a Hausdorff distance of 1.5 of another play/attack/trajectory matrix. Thus, distance can essentially be thought of as similarity in terms of the degree to which an attack is similar to another. This suggests that at the match level, a distance threshold of 4.0 is more appropriate to ensure an adequate number of SSD sub-matrices within attacks can be obtained. Comparing the "No. of distinct matches" column with the SSD attacks by the "No. matches" column in Table 2 shows that most of the teams' matches within the season contained some SSD attack(s). While the analysis in Table 2 is useful for selecting appropriate parameters, note that it does not show the number of attacks with SSD sub-matrices within each match. It should also be noted that there is overlap in the attack and match counts in Table 2, e.g., sub-matrices and attacks/matches with a minimum length of 10 can also be obtained with a minimum length of 5 and 8.

The experiments using the MA-Stat-DSM algorithm were run on an Intel® Xeon® CPU E5-2697 v2 @ 2.70GHz Linux CentOS server machine (Linux version 3.10.0), running at CPU 3.2 GHz, using 128 GB of RAM. The most influential factor affecting the run time of the MA-Stat-DSM algorithm was the distance threshold parameter, taking nearly five times as long on average (per MA-Stat-DSM iteration) to run the algorithm with a distance threshold of 4 compared to a distance threshold of 1.5. A distance threshold of 20 was not feasible on our dataset because of the computational complexity, which was found to be prohibitive when running MA-Stat-DSM on a particular team's attacks in a specific match.

Tab. 1: MA-Stat-DSM parameters (top) and data preprocessing parameters (bottom)

Minimum length L	**5, 8, 10**
Distance threshold ε	**1.5, 4**
Number of permutations B	**1000**
Significance level α	**0.05**
Set of agents K	**{ball, shooter, last passer, shooter defender, last passer defender}**
Time interval	**t2 to t0 (as per Figure 1)**

5.2 Visualization and interpretation of SSD sub-matrix examples

In this subsection, we provide some SSD sub-matrix result examples that were obtained when applying MA-Stat-DSM to team attacks in selected matches (listed in Table 3) involving the two top teams and

Tab. 2: Number of matches containing attacks with statistically significantly discriminative (SSD) sub-matrices, and the number of attacks containing SSD sub-matrices, for $\varepsilon = 1.5$ and $\varepsilon = 4$, and $L = 5, L = 8, L = 10$ for the two top and two bottom teams' matches in the 2015/16 season.

team	No. matches	Distance threshold (ε)	No. SSD attacks	No. of distinct matches containing attacks with SSD sub-matrices	No. of attacks with SSD sub-matrices ($L = 5$)	No. of attacks with SSD sub-matrices ($L = 8$)	No. of attacks with SSD sub-matrices ($L = 10$)	No. of distinct matches containing attacks with SSD sub-matrices ($L = 5$)	No. of distinct matches containing attacks with SSD sub-matrices ($L = 8$)	No. of distinct matches containing attacks with SSD sub-matrices ($L = 10$)
GSW	40	4	625	40	242	198	185	33	30	31
		1.5	4	4	2	1	1	2	1	1
PHI	40	4	815	35	306	270	239	34	33	31
		1.5	4	4	2	1	1	2	1	1
LAL	41	4	786	31	294	255	237	31	29	28
		1.5	10	4	6	2	2	4	1	1
CLE	36	4	692	30	244	229	219	28	28	27
		1.5	2	2	2	0	0	2	0	0

Tab. 3: Number of attacking plays in the match and the number of those that contained SSD sub-matrices for selected matches involving the top- and bottom-performing teams in the 2015/16 NBA season (with $\varepsilon = 4$, $L = 5$).

Match Date	Home Team	Away Team	No. of attacks by the Home Team	No. of attacks by the Away Team	No. of Home Team attacks with SSD sub-matrices	No. of Away Team attacks with SSD sub-matrices
10-Jan-16	PHI	CLE	17	20	4	2
05-Jan-16	LAL	GSW	20	32	14	17
25-Dec-15	GSW	CLE	20	28	1	23
20-Dec-15	CLE	PHI	20	24	2	8
01-Dec-15	PHI	LAL	23	19	11	3

two bottom teams in the 2015/16 NBA season. Table 2 shows that a minimum length parameter of 5 obtained more SSD attacks within matches than $L = 8$ or $L = 10$. Furthermore, as mentioned previously, a distance threshold of 4 is preferred when applying MA-Stat-DSM at the match level. The remainder of the parameters for the following results are $B = 1000$, $\alpha = 0.05$, $K = \{$ball, shooter, shooter defender, last passer, last passer defender$\}$, and with the time interval of t2 to t0 (Figure 1).

The discriminative portions of the attacks (the SSD sub-matrices) in Figures 4 to 11 are denoted with plus signs and the remainder of the trajectories are also shown as per the time interval t2 to t0 (Figure 1), which is the time interval from when the last passer received the ball to when a shot is made or the ball is turned over. The movement of the agents on the court in these figures is indicated by the colour's progression from light to dark. Recall that the SSD sub-matrix consists of the sub-trajectories of each of the agents at the same timestamps (i.e., the columns are temporally aligned), and it represents the portion of the attack that discriminates between the effective and ineffective labeled attacks (represented by trajectory matrices) for a particular team in a specific particular match.

An SSD sub-matrix from the 5 January 2016 match between the Golden State Warriors and the Los Angeles Lakers, which was obtained with MA-Stat-DSM applied to Golden State's attacks in this match, is displayed in Figure 4. As shown in Table 3, 17 Golden State attacks containing SSD sub-matrices were obtained in this particular match with the above-mentioned parameters, of which Figure 4 is one. In this Golden State attack against the Lakers, the points comprising the discriminative sub-trajectory of the Golden State shooter are much more spread out relative to the slower movement made in response by the shooter defender. This rapid movement enabled the shooter to reach the edge of the three-point arc to subsequently be in a position to make a 3-point shot, which led to the attack being effective (the shot was also ultimately successful).

An SSD sub-matrix in a Lakers attack, obtained by applying MA-Stat-DSM to the same 5 January 2016 match between the Golden State Warriors and the Los Angeles Lakers, with the parameters mentioned, is shown in Figure 5. As shown in Table 3, 14 Lakers attacks containing SSD sub-matrices were obtained in this particular match, of which Figure 5 is one. In this play, the SSD portion of the play shows the movement of the Lakers shooter with the ball. The speed of the movement of the drive forward by the Lakers shooter slightly outpacing the defensive movement in the same direction of the shooter defender, thereby rendering the attack effective (although a shot attempt was made, the shot itself was ultimately unsuccessful).

In both of the preceding examples, player movement — off-the-ball movement of the shooter in the first case, and on-the-ball movement in the second example — appeared to be the primary factor that resulted in the attacks being effective. The movement of the ball through passing is another key factor that determines whether a play is ultimately effective or ineffective. Figure 6 shows an ineffective Golden State

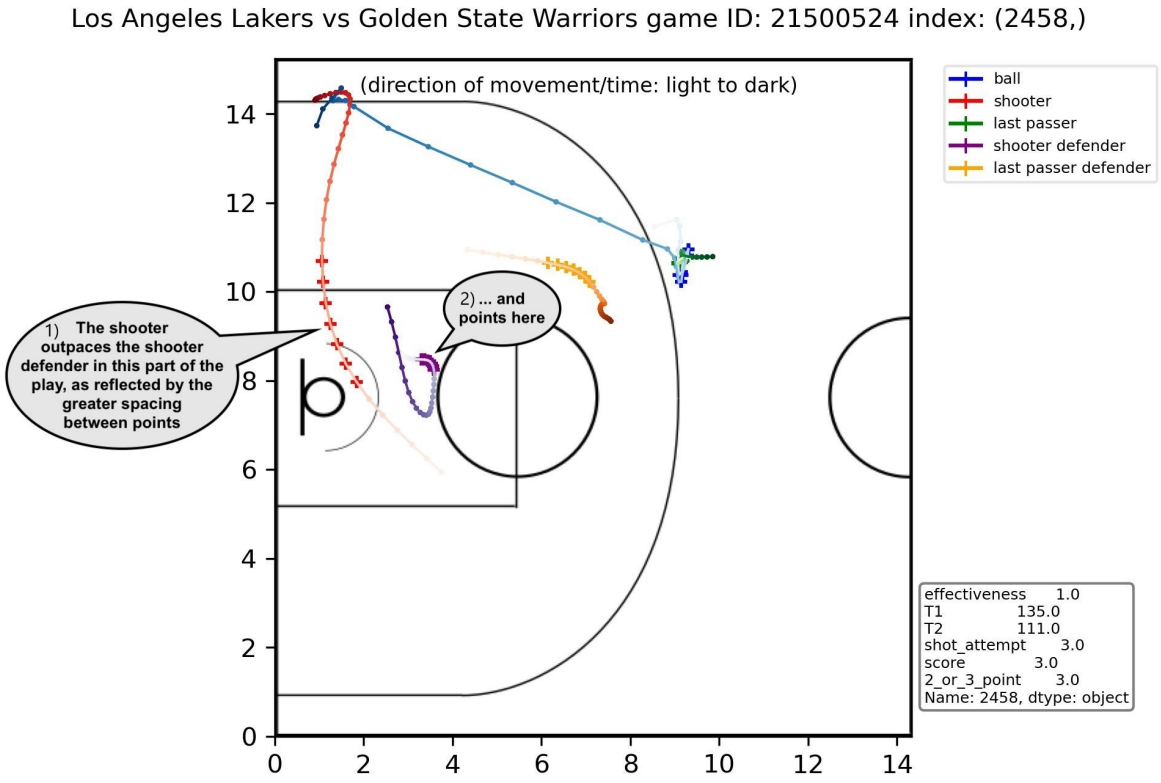


Fig. 4: An effective Golden State attack with an SSD sub-matrix result from the 5 January 2016 match between the Golden State Warriors and Los Angeles Lakers. The agent sub-trajectories constituting the SSD sub-matrix are denoted by plus signs. Time progression is displayed by the trajectories' colour transitioning from light to dark.

attack, from the same 5 January 2016 against the Lakers, in which the SSD sub-matrix covers most of the agent movements within the T2 interval. In this play, the pass made by the last passer while moving in a backward direction results in a pass that appears to be covered by the shooter defender, rendering the attack ineffective (a shot was attempted but was unsuccessful).

Figure 7 shows an ineffective Lakers attack SSD sub-matrix from the same match against Golden State in which the pass from open space from the last passer to the shooter does not constitute any part of the discriminative portion of the attack. The position of the shooter at the time they receive the pass appears to be unfavourable, however, and is covered by the shooter defender. At the same time, the last passer was making a rapid run off the ball near the free-throw line, and it may have been a better option to pass the ball back to them, e.g., to make a lay-up, rather than attempting a 3-point shot outside the three-point arc under defensive pressure, which was ultimately unsuccessful.

In the discriminative portion of the Lakers attack shown in Figure 8, from a match against the Philadelphia 76ers on 1 December 2015, the Laker's last passer makes a pass with roughly a 45-degree angle to the free-throw line while the shooter simultaneously makes a run perpendicular to the free-throw lane line to receive the ball from the last passer, whose pass managed to avoid the last passer defender despite the last passer defender closely tracking the last passer. During this time, the discriminative portion of the shooter defender was relatively static, and although the play was effective, the shot attempt was unsuccessful.

The last passer's pass to the shooter does not comprise any part of the discriminative portion of the Cleveland attack shown in Figure 9, which is from a match played on the 10 January 2016 between the

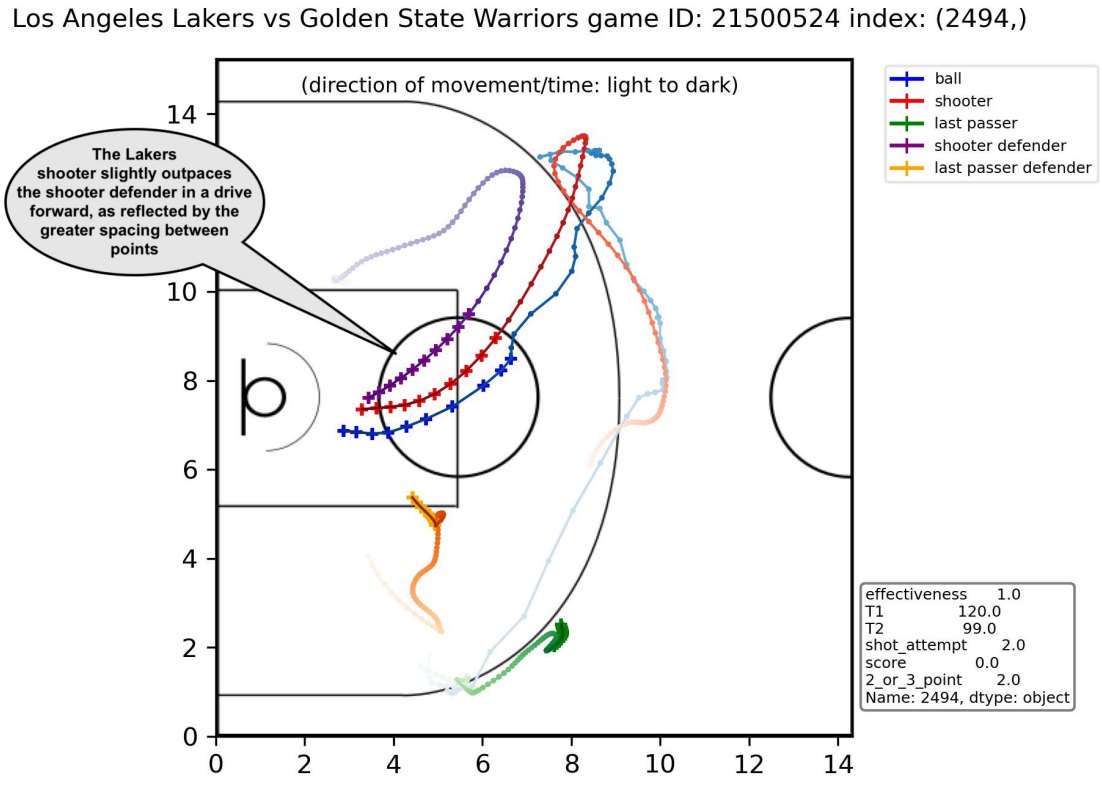


Fig. 5: An effective Lakers attack with an SSD sub-matrix from the 5 January 2016 match between the Golden State Warriors and Los Angeles Lakers. The agent sub-trajectories constituting the SSD sub-matrix are denoted by plus signs. Time progression is displayed by the trajectories' colour transitioning from light to dark.

Cleveland Cavaliers and Philadelphia 76ers. The discriminative portion that rendered the play ineffective was the more rapid speed of the movement of the shooter defender towards the shooter (relative to the speed of the shooter's movement), who had moved backwards to the edge of the three-point arc after receiving the pass to attempt an unsuccessful 3 pointer.

Similarly, the pass did not form any part of the discriminative sub-matrix in the ineffective 76ers attack shown in Figure 10, which is from a match played on 20 December 2015 between the Cleveland Cavaliers and Philadelphia 76ers. The most interesting sub-trajectories in this attack seem to be those of the last passer and last passer defender, the latter tracking the movement of the last passer on the inside into the free-throw lane, perhaps meaning that the shooter could not get a pass away to them and therefore instead attempted a shot that missed from just outside the in-the-paint area.

Similar to the attack in Figure 6, sometimes the discriminative sub-matrix formed all (or nearly all) of the trajectory matrix during the time interval from t2 to t0. In Figure 11, the trajectory sub-matrix and matrix were the same, indicating that the whole attack was relevant in discriminating between, in this case, the effective and ineffective attacks of Cleveland in their 25 December match against Golden State. This particular attack ended in a successful 3-point shot by the Cleveland shooter, although the distance between the shooter and shooter defender by the time the shot was made meant the play was labelled as ineffective.

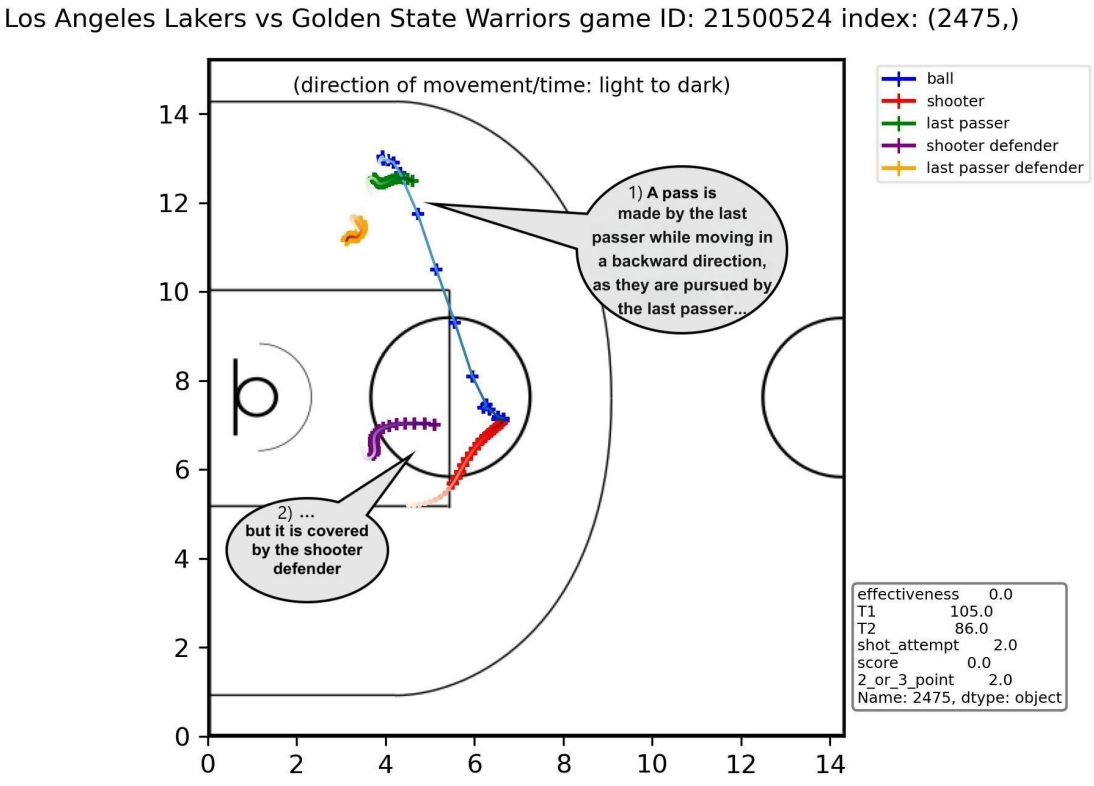


Fig. 6: An ineffective Golden State attack SSD sub-matrix from the 5 January 2016 match between the Golden State Warriors and Los Angeles Lakers. The agent sub-trajectories constituting the SSD sub-matrix are denoted by plus signs. Time progression is displayed by the trajectories' colour transitioning from light to dark.

6 Discussion

In this paper, a multi-agent statistically discriminative sub-trajectory mining (MA-Stat-DSM) method was proposed, which extends the (single-agent) Stat-DSM [23] method to the multi-agent setting. MA-Stat-DSM was applied to Stats Perform SportVU NBA trajectory data to identify the most relevant parts of attacking plays that discriminate between effective and ineffective attacks, with attack effectiveness defined based on the concept of wide-open shots, shooter positions and distances from the closest defender, and historical 3-point shot success probabilities. Distance between respective plays and parts of plays was computed using Hausdorff distance. The benefits of applying MA-Stat-DSM to labeled sets of agent trajectories as opposed to simply interpreting visualized trajectories that are labeled as effective/ineffective is that the method can automatically identify the part of the play that was most relevant in terms of which agents were involved (e.g., whether a pass was involved), and whether on-the-ball or off-the-ball movements (or both) were most relevant.

The number of ineffective or effective attacks identified by MA-Stat-DSM depends on parameters of the algorithm, especially the distance threshold. Increasing the distance threshold will generally result in a larger number of SSD attacks being identified by MA-Stat-DSM. The minimum length parameter, however, did not appear to have any particular effect on the number of SSD attacks obtained.

The MA-Stat-DSM method was demonstrated by applying it to labeled team attacks in a match containing the trajectories of five agents (the ball, two attacking players, and two defending players), in the time interval from when the last passer receives the ball until a shot is attempted (or the ball is turned over).

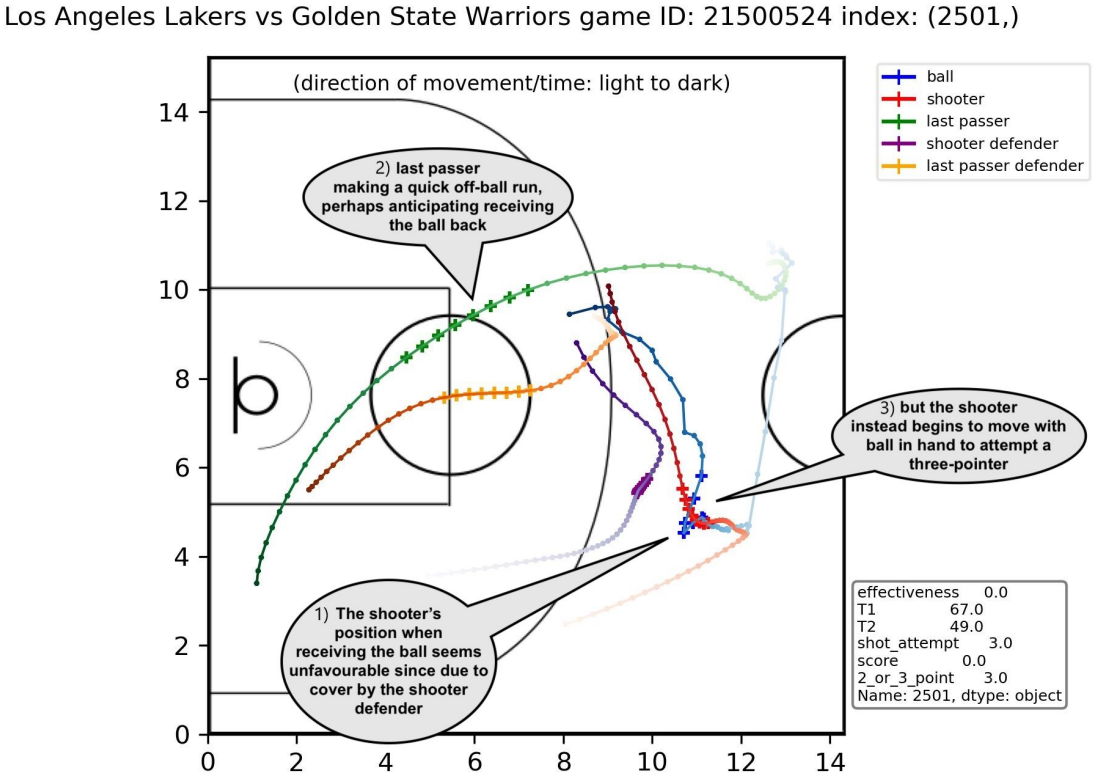


Fig. 7: An ineffective Lakers attack SSD sub-matrix from the 5 January 2016 match between the Golden State Warriors and Los Angeles Lakers. The agent sub-trajectories constituting the SSD sub-matrix are denoted by plus signs. Time progression is displayed by the trajectories' colour transitioning from light to dark.

Matches involving the two top-performing teams from the 2015/16 NBA season, Cleveland and Golden State, and/or the bottom teams in the Eastern and Western conferences (the 76ers and Lakers, respectively) were considered.

In some cases, the obtained discriminative sub-matrix encompassed all (or nearly all) of the trajectory matrix (e.g., Figure 11). In some attacks, the discriminative portion included a pass within it, sometimes it involved player movement only, i.e., the pass to the shooter had already been completed, and sometimes included relevant off-the-ball movements (e.g., Figure 4), and sometimes included both a pass and player movements (e.g., Figure 6).

The obtained results suggested that the distance threshold is the key parameter of MA-Stat-DSM in determining how many statistically significantly discriminative (SSD) sub-matrices are obtained by the algorithm. On the datasets to which MA-Stat-DSM was applied in the current study, the attacks of a specific team in a particular match, a distance threshold of 4.0 was found to be more appropriate than a distance threshold of 1.5 to obtain an adequate number of SSD sub-matrices. Even when the MA-Stat-DSM was iterated over all of a team's matches in the entire season, a distance threshold of 1.5 only obtained a very small number of SSD sub-matrices, suggesting that a distance threshold somewhere between 1.5 and 4.0 may be appropriate if conducting a season-level analysis to identify the most important parts of attacks across a team's whole season. The distance threshold was also the most important parameter in determining the run time of MA-Stat-DSM, with a distance threshold of 4.0 taking nearly five times as long to run per iteration (a team's attacks in a single match) compared to a distance threshold of 1.5.

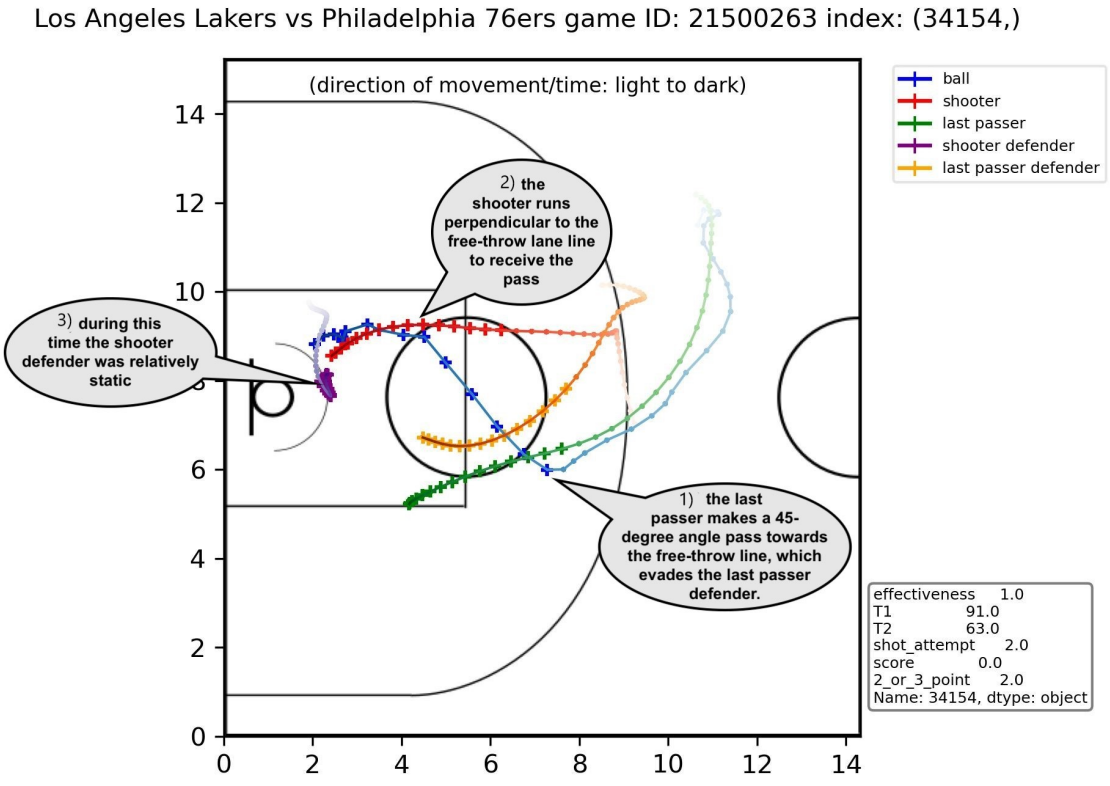


Fig. 8: An effective Lakers attack SSD sub-matrix from the 1 December 2015 match between the Philadelphia 76ers and Los Angeles Lakers. The agent sub-trajectories constituting the SSD sub-matrix are denoted by plus signs. Time progression is displayed by the trajectories' colour transitioning from light to dark.

There are potential benefits in utilizing MA-Stat-DSM to identify the most relevant parts of attacks rather than watching and coding video, which is time-consuming for coaches and analysts and may result in biases in the types of play deemed most relevant. MA-Stat-DSM can automatically identify aspects of multi-agent behaviour through plays, or parts of plays, which may not be obvious to coaches or analysts. As previously mentioned, MA-Stat-DSM, unlike machine learning methods, does not require complex feature engineering, e.g., of changes in position, player speeds, accelerations, etc. Furthermore, the proposed method is more intuitive than deep learning methods, which are generally black-box.

In this study, the set of MA-Stat-DSM parameters was restricted to reduce the number of possible permutations. In future research, it could be confirmed whether increasing the distance threshold and the significance level (e.g., from 0.05 to 0.1) would increase the number of discriminative sub-matrices obtained. Other base distances for Hausdorff distance other than Euclidean distance could also be trialed. Rather than pre-defined areas of the court being used to compute player shooting success probabilities, a more sophisticated approach by partitioning the court could also be performed using classification trees as per [51]. A limitation of MA-Stat-DSM is that its computational complexity meant it could not be applied to datasets with a large number of trajectories, e.g., a team's attacks from an entire season. Thus, we considered a team's set of attacks from a single match, and iterated MA-Stat-DSM over each of the team's matches in the season. Future work could improve the efficiency of MA-Stat-DSM. In some cases (e.g., the pass in Figure 8), it would have been useful to have the z-coordinate of the ball. Although we considered five agents (four players) in this study, e.g., for reasons previously mentioned in subsection 4.1, knowing the positions of other players may in fact have aided interpretation in some cases (e.g., the attack in Figure 4

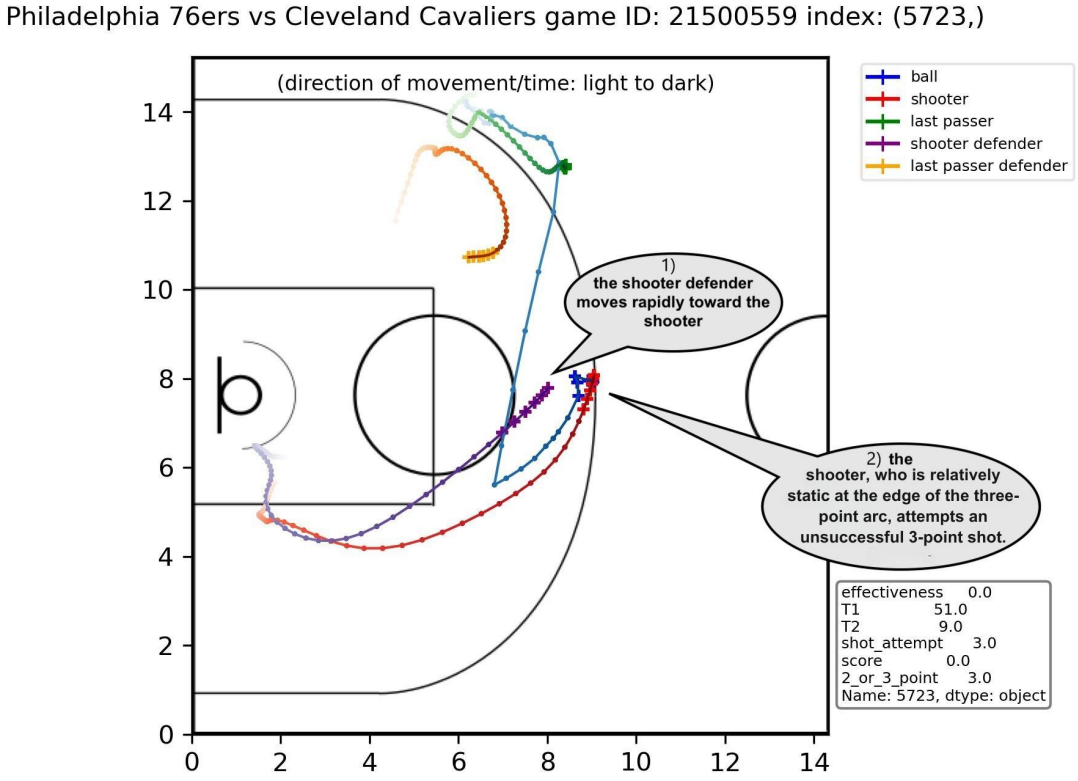


Fig. 9: An ineffective Cleveland attack SSD sub-matrix from the 10 January 2016 match between the Cleveland Cavaliers and Philadelphia 76ers. The agent sub-trajectories constituting the SSD sub-matrix are denoted by plus signs. Time progression is displayed by the trajectories' colour transitioning from light to dark.

may have involved a ball screen). With an appropriate label defined, MA-Stat-DSM could be applied to multi-agent trajectories from other sports and other domains. Finally, the method could be generalized, e.g., such that discriminative sub-matrices need not always have a fixed number of agents.

References

[1] Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science 7*(1), 131–153.

[2] Ai, S., J. Na, V. De Silva, and M. Caine (2021). A novel methodology for automating spatio-temporal data classification in basketball using active learning. In *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*, pp. 39–45. IEEE.

[3] Alcorn, M. A. and A. Nguyen (2021). baller2vec: A multi-entity transformer for multi-agent spatiotemporal modeling. *arXiv preprint arXiv:2102.03291*.

[4] Bunker, R., K. Fujii, H. Hanada, and I. Takeuchi (2021). Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: an application to rugby union. *PLOS One 16*(9), e0256329.

[5] Cao, Y., J. Zhu, and F. Gao (2016). An algorithm for mining moving flock patterns from pedestrian trajectories. In *Web Technologies and Applications: APWeb 2016 Workshops, WDMA, GAP, and SDMA, Suzhou, China, September 23-25, 2016, Proceedings*, pp. 310–321. Springer.

[6] Carling, C., J. Bloomfield, L. Nelsen, and T. Reilly (2008). The role of motion analysis in elite soccer: contemporary performance measurement techniques and work rate data. *Sports Medicine 38*, 839–862.

[7] Chen, C.-H., T.-L. Liu, Y.-S. Wang, H.-K. Chu, N. C. Tang, and H.-Y. M. Liao (2015). Spatio-temporal learning of basketball offensive strategies. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1123–1126.

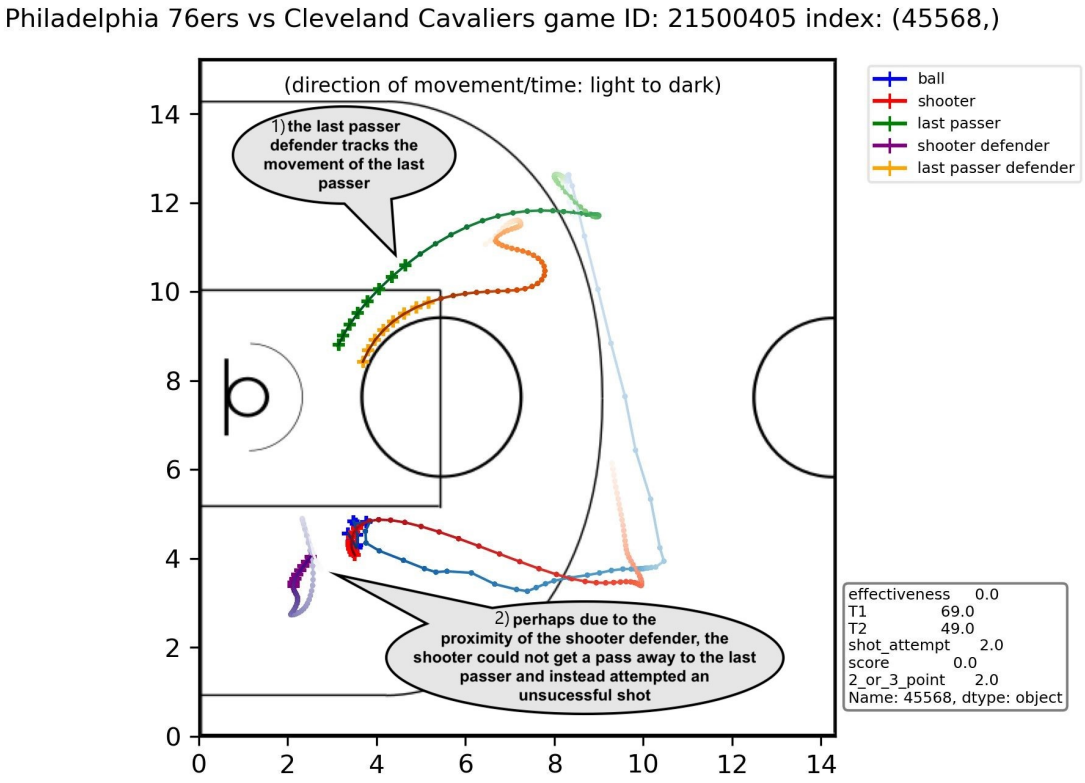


Fig. 10: An ineffective 76ers attack SSD sub-matrix from the 20 December 2015 match between the Cleveland Cavaliers and Philadelphia 76ers. The agent sub-trajectories constituting the SSD sub-matrix are denoted by plus signs. Time progression is displayed by the trajectories' colour transitioning from light to dark.

[8] Facchinetti, T., R. Metulini, and P. Zuccolotto (2023). Filtering active moments in basketball games using data from players tracking systems. *Annals of Operations Research 325*(1), 521–538.

[9] Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society 85*(1), 87–94.

[10] Fujii, K., Y. Inaba, and Y. Kawahara (2017). Koopman spectral kernels for comparing complex dynamics: Application to multiagent sport plays. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD'17)*, pp. 127–139. Springer.

[11] Fujii, K., T. Kawasaki, Y. Inaba, and Y. Kawahara (2018). Prediction and classification in equation-free collective motion dynamics. *PLoS Computational Biology 14*(11), e1006545.

[12] Fujii, K., N. Takeishi, Y. Kawahara, and K. Takeda (2020). Policy learning with partial observation and mechanical constraints for multi-person modeling. *arXiv preprint arXiv:2007.03155*.

[13] Fujii, K., K. Yokoyama, T. Koyama, A. Rikukawa, H. Yamada, and Y. Yamamoto (2016). Resilient help to switch and overlap hierarchical subsystems in a small human group. *Scientific Reports 6*, 23911.

[14] Goes, F., L. Meerhoff, M. Bueno, D. Rodrigues, F. Moura, M. Brink, M. Elferink-Gemser, A. Knobbe, S. Cunha, R. Torres, et al. (2021). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science 21*(4), 481–496.

[15] Gudmundsson, J. and M. Horton (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys (CSUR) 50*(2), 1–34.

[16] Herrera, F., S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, and S. Vluymans (2016). *Multiple Instance Learning: Foundations and Algorithms*. Springer.

[17] Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural Computation 9*(8), 1735–1780.

[18] Hrovat, G., I. Fister Jr, K. Yermak, G. Stiglic, and I. Fister (2015). Interestingness measure for mining sequential patterns in sports. *Journal of Intelligent & Fuzzy Systems 29*(5), 1981–1994.

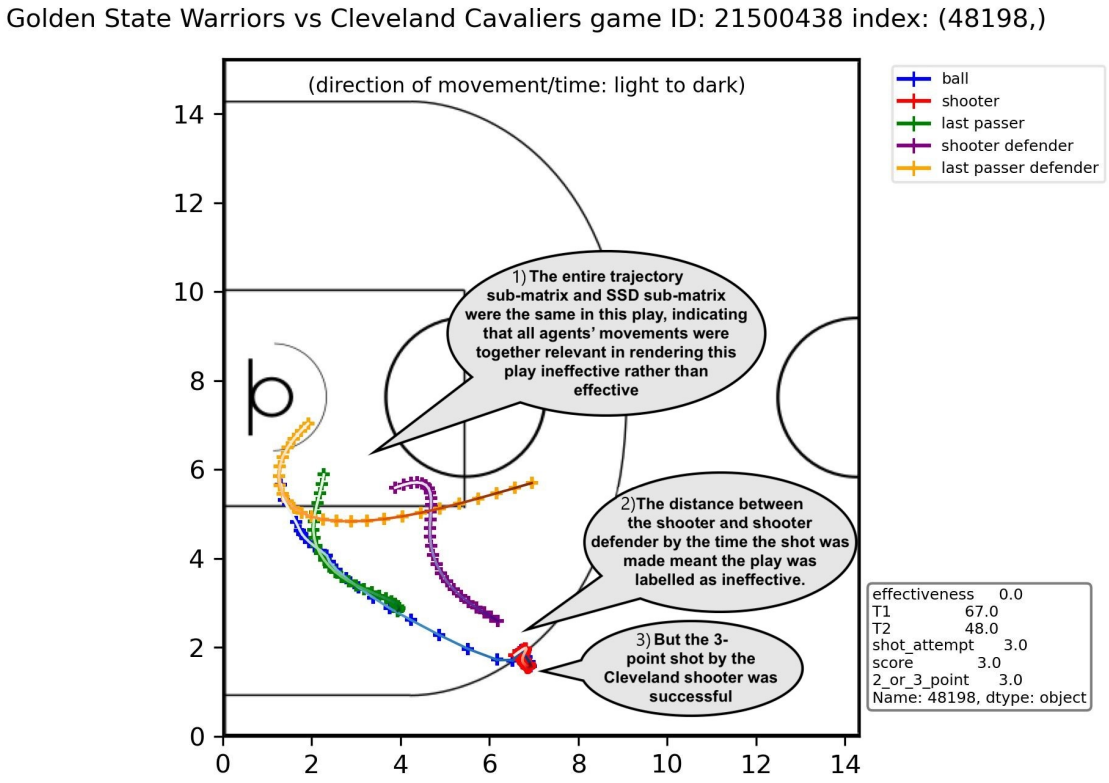


Fig. 11: An ineffective (but 3-point scoring) Cleveland attack SSD sub-matrix from the 25 December 2015 match between the Cleveland Cavaliers and Golden State Warriors. The agent sub-trajectories constituting the SSD sub-matrix are denoted by plus signs. Time progression is displayed by the trajectories' colour transitioning from light to dark.

[19] Hughes, M. D. and R. M. Bartlett (2002). The use of performance indicators in performance analysis. *Journal of Sports Sciences 20*(10), 739–754.

[20] Huttenlocher, D. P., G. A. Klanderman, and W. J. Rucklidge (1993). Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence 15*(9), 850–863.

[21] Karcher, C. and M. Buchheit (2014). On-court demands of elite handball, with special reference to playing positions. *Sports Medicine 44*, 797–814.

[22] Le, H. M., Y. Yue, P. Carr, and P. Lucey (2017). Coordinated multi-agent imitation learning. In *International Conference on Machine Learning*, pp. 1995–2003. PMLR.

[23] Le Vo, D. N., T. Sakuma, T. Ishiyama, H. Toda, K. Arai, M. Karasuyama, Y. Okubo, M. Sunaga, H. Hanada, Y. Tabei, et al. (2020). Stat-DSM: Statistically discriminative sub-trajectory mining with multiple testing correction. *IEEE Transactions on Knowledge and Data Engineering*.

[24] Li, Z., J. Han, M. Ji, L.-A. Tang, Y. Yu, B. Ding, J.-G. Lee, and R. Kays (2011). Movemine: Mining moving object data for discovery of animal movement patterns. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*(4), 1–32.

[25] Lucey, P., A. Bialkowski, P. Carr, Y. Yue, and I. Matthews (2014). How to get an open shot: Analyzing team movement in basketball using tracking data. In *Proceedings of the 8th Annual MIT SLOAN Sports Analytics Conference*.

[26] Mazimpaka, J. D. and S. Timpf (2016). Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science 2016*(13), 61–99.

[27] McIntyre, A., J. Brooks, J. Guttag, and J. Wiens (2016). Recognizing and analyzing ball screen defense in the NBA. In *Proceedings of the MIT Sloan Sports Analytics Conference, Boston, MA, USA*, pp. 11–12.

[28] McQueen, A., J. Wiens, and J. Guttag (2014). Automatically recognizing on-ball screens. In *2014 MIT Sloan Sports Analytics Conference*.

[29] Metulini, R., M. Manisera, and P. Zuccolotto (2018). Modelling the dynamic pattern of surface area in basketball and its effects on team performance. *Journal of Quantitative Analysis in Sports 14*(3), 117–130.

[30] Papadimitriou, C. H. and K. Steiglitz (1982). *Combinatorial optimization*, Volume 24. Prentice Hall Englewood Cliffs.
[31] Papalexakis, E. and K. Pelechrinis (2018). thoops: A multi-aspect analytical framework for spatio-temporal basketball data. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 2223–2232.
[32] Raabe, D., R. Nabben, and D. Memmert (2023). Graph representations for the analysis of multi-agent spatiotemporal sports data. *Applied Intelligence 53*(4), 3783–3803.
[33] Rein, R. and D. Memmert (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus 5*(1), 1–13.
[34] Sampaio, J., T. McGarry, J. Calleja-González, S. Jiménez Sáiz, X. Schelling i del Alcázar, and M. Balciunas (2015). Exploring game performance in the National Basketball Association using player tracking data. *PLOS One 10*(7), e0132894.
[35] Shah, R. and R. Romijnders (2016). Applying deep learning to basketball trajectories. *arXiv preprint arXiv:1608.03793*.
[36] Sicilia, A., K. Pelechrinis, and K. Goldsberry (2019). Deephoops: Evaluating micro-actions in basketball using deep feature representations of spatio-temporal data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2096–2104.
[37] Skinner, B. and S. J. Guy (2015). A method for using player tracking data in basketball to learn player skills and predict team performance. *PLOS One 10*(9), e0136393.
[38] Su, H., S. Liu, B. Zheng, X. Zhou, and K. Zheng (2020). A survey of trajectory distance measures and performance evaluation. *The VLDB Journal 29*(1), 3–32.
[39] Taha, A. A. and A. Hanbury (2015). An efficient algorithm for calculating the exact Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(11), 2153–2163.
[40] Terada, A., K. Tsuda, and J. Sese (2013). Fast Westfall-Young permutation procedure for combinatorial regulation discovery. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 153–158. IEEE.
[41] Terner, Z. and A. Franks (2021). Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application 8*, 1–23.
[42] Tian, C., V. De Silva, M. Caine, and S. Swanson (2019). Use of machine learning to automate the identification of basketball strategies using whole team player tracking data. *Applied Sciences 10*(1), 24.
[43] Wang, K.-C. and R. Zemel (2016). Classifying NBA offensive plays using neural networks. In *Proceedings of MIT Sloan Sports Analytics Conference*, Volume 4.
[44] Westfall, P. H. and S. S. Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Volume 279. John Wiley & Sons.
[45] Yoon, Y., H. Hwang, Y. Choi, M. Joo, H. Oh, I. Park, K.-H. Lee, and J.-H. Hwang (2019). Analyzing basketball movements and pass relationships using realtime object tracking techniques based on deep learning. *IEEE Access 7*, 56564–56576.
[46] Zhao, Y., S. Shang, Y. Wang, B. Zheng, Q. V. H. Nguyen, and K. Zheng (2018). Rest: A reference-based framework for spatio-temporal trajectory compression. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2797–2806.
[47] Zhao, Y., R. Yang, G. Chevalier, R. C. Shah, and R. Romijnders (2018). Applying deep bidirectional lstm and mixture density network for basketball trajectory prediction. *Optik 158*, 266–272.
[48] Zheng, Y. (2015). Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST) 6*(3), 1–41.
[49] Ziyi, Z., R. Bunker, K. Takeda, and K. Fujii (2023). Multi-agent deep-learning based comparative analysis of team sport trajectories. *IEEE Access*.
[50] Ziyi, Z., K. Takeda, and K. Fujii (2022). Cooperative play classification in team sports via semi-supervised learning. *International Journal of Computer Science in Sport 21*(1), 111–121.
[51] Zuccolotto, P., M. Sandri, and M. Manisera (2021). Spatial performance indicators and graphs in basketball. *Social Indicators Research 156*, 725–738.

Appendix

Code

The MA-Stat-DSM code is available on GitHub: https://github.com/rorybunker/ma-stat-dsm

Statistical testing & pruning properties of (MA)-Stat-DSM

The statistical testing and pruning properties of MA-Stat-DSM are essentially analogous to the original Stat-DSM [23]. Therefore, in this subsection (and in Figure 2 in the main text), we provide a brief explanation of the statistical testing and pruning properties of (MA-)Stat-DSM and refer the reader to [23] for full details.

Stat-DSM (MA-Stat-DSM) represents sub-trajectories (sub-matrices) in the form of a tree, which is pruned to remove sub-trajectories (sub-matrices) that are guaranteed to not be discriminative (this pruning criterion is shown in line 18 of the MA-Stat-DSM algorithm pseudo-code in Algorithm 1). Stat-DSM (MA-Stat-DSM) uses Fisher's Exact Test (FET) [1, 9] to determine the statistical significance of a sub-trajectory (sub-matrix) using a contingency table with the number of trajectories (matrices) that, respectively, contain and do not contain sub-trajectories (sub-matrices) within a distance of ε. A correction for multiple-testing bias is also incorporated, which is necessary due to the calculation of p-values for a large number of trajectories (trajectory matrices), and is conducted using the Westfall-Young (WY) method [40, 44]. Sub-trajectories (sub-matrices) are only identified as SSD if their p-value is less than their adjusted significance level, δ, which is, in turn, less than $\alpha = 0.05$. The dataset labels are permuted $B = 1000$ times as part of this procedure. The pruning and WY methods are applied simultaneously to reduce complexity (Step 1, Figure 2).

Funding: This work was supported by JSPS KAKENHI (Grant Numbers 19H04941 and 20H04075) and JST PRESTO (JPMJPR20CA).