INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL



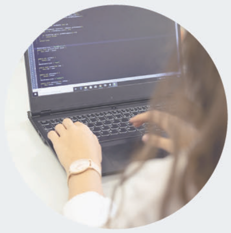

Modelos Híbridos para Previsão de Resultados de Jogos da Premier League Usando Machine Learning e Análise de Sentimento

RUBENS FABRÍCIO DO ROSÁRIO SOARES NASCIMENTO
Outubro de 2025

P.PORTO

INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL



Hybrid Models for Predicting Premier League Match Outcomes Using Machine Learning and Sentiment Analysis

RUBENS FABRÍCIO DO ROSÁRIO SOARES NASCIMENTO
Outubro de 2025

P.PORTO

Hybrid Models for Predicting Premier League Match Outcomes

Using Machine Learning and Sentiment Analysis

Rubens Fabrício do R. S. Nascimento

Student No.: 1222750

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science, Specialisation Area of Artificial Intelligence

Supervisor: Carlos Fernando da Silva Ramos, Full Professor, Institute of Engineering - Polytechnic of Porto (ISEP/IPP)

Evaluation Committee:

President:

Isabel Cecília Correia da Silva Praça Gomes Pereira, Associate Professor, Institute of Engineering - Polytechnic of Porto (ISEP/IPP)

Members:

Diogo Emanuel Pereira Martinho, Assistant Professor, Institute of Engineering - Polytechnic of Porto (ISEP/IPP)

Carlos Fernando da Silva Ramos, Full Professor, Institute of Engineering - Polytechnic of Porto (ISEP/IPP)

Porto, October 2025

Abstract

This study explores whether combining structured match statistics with pre-match tweet sentiment can enhance probabilistic forecasting of football results. Focusing on English Premier League fixtures, it aligns social signals with each game and compares three families of models: those based solely on statistics, those relying only on tweets, and hybrid approaches that integrate both. The evaluation respects the chronological order of matches, employing sequential training and validation together with a strict 2024/25 holdout. In terms of assessment, Log Loss serves as the primary metric, complemented by calibration measures (ECE, Brier, RPS) as well as accuracy.

When comparing different families of models, statistical learners provide the strongest foundation. Within this group, an RBF-SVM delivers a holdout Log Loss of 0.9066 with 58.16% accuracy, while a regularised Logistic Regression remains competitive, suggesting that engineered features capture a substantial linear signal. By contrast, tweet-only models offer useful but weaker contributions. The best-performing configuration, a Linear SVM applied to SBERT-MPNet embeddings, records a Log Loss of 1.0313 and an accuracy of 47.89%, yet generalises consistently across both validation and test.

Across the different model families, hybrid approaches provide the most consistent improvements. In particular, Early Fusion with Logistic Regression, which combines sentiment with structured inputs, delivers 59.74% accuracy and a Log Loss of 0.8954 on the holdout, together with a Brier Score of 0.1758 and an RPS of 0.1171. Moreover, Residual Stacking extends these gains by further reducing both Log Loss and Expected Calibration Error compared with the statistical baseline, with the benefits especially clear in lower-confidence fixtures and in predicting draws.

The main improvements come from modest probability refinements that reduce error penalties without frequent class flips, while also enhancing calibration. At the same time, certain limitations remain, including the focus on a single league, the risk of temporal drift in team performance, and the presence of noise, ambiguity, and attention bias in social text. Taken together, the findings demonstrate that combining structured match data with curated sentiment yields robust and well-calibrated forecasts, particularly valuable in uncertain fixtures and in outcomes that are traditionally harder to predict.

Keywords: Football match prediction, English Premier League, Machine Learning, Linear Models, Kernel Methods, Tree-based Methods, Neural Networks, Sentiment Analysis, Ensemble Methods, Hybrid Models

Resumo

Este estudo explora a possibilidade da combinação de estatísticas dos jogos com o sentimento expresso em tweets publicados antes das partidas pode melhorar a previsão probabilística de resultados de futebol. Com foco em jogos da English Premier League, o trabalho alinha os sinais sociais a cada encontro e compara três famílias de modelos: os baseados apenas em estatísticas, os que recorrem exclusivamente a tweets e as abordagens híbridas que integram ambas as fontes. A avaliação respeita a ordem cronológica dos jogos, recorrendo a treino e validação sequenciais, bem como a um holdout rigoroso correspondente à época 2024/2025. Para a medição de desempenho, utiliza-se o Log Loss como métrica principal, complementado por medidas de calibração (ECE, Brier, RPS) e pela acurácia.

Ao comparar diferentes famílias de modelos, os baseados em estatísticas fornecem a base mais sólida. Entre estes, um RBF-SVM alcança no holdout um Log Loss de 0.9066 com 58.16% de acurácia, enquanto uma Regressão Logística regularizada permanece competitiva, sugerindo que as features projectadas captam um sinal linear relevante. Em contraste, os modelos baseados apenas em tweets oferecem contributos úteis, mas mais modestos. O melhor resultado deste grupo, obtido com um SVM Linear aplicado a embeddings SBERT-MPNet, regista um Log Loss de 1.0313 e uma acurácia de 47.89%, demonstrando ainda assim consistência entre validação e teste.

Entre as diferentes famílias de modelos, as abordagens híbridas proporcionam os ganhos mais consistentes. Em particular, a fusão antecipada com Regressão Logística, que integra sentimento com informação estatística, atinge 59.74% de acurácia e um Log Loss de 0.8954 no holdout, acompanhados por um Brier Score de 0.1758 e um RPS de 0.1171. Além disso, o Residual Stacking reforça estes ganhos ao reduzir ainda mais, tanto o Log Loss como o Expected Calibration Error face ao modelo estatístico de base, com benefícios especialmente claros em jogos de maior incerteza e na previsão de empates.

As principais melhorias resultam de ajustes subtis nas probabilidades, que reduzem penalizações de erro sem alterar frequentemente a classe prevista, ao mesmo tempo que reforçam a calibração. Persistem, contudo, algumas limitações: o foco num único campeonato, o risco de desvio temporal no desempenho das equipas e a presença de ruído, ambiguidades e viés de atenção nos tweets. Em síntese, os resultados mostram que a combinação de dados estatísticos dos jogos com sentimento extraído de redes sociais produz previsões robustas e bem calibradas, particularmente valiosas em jogos incertos e em resultados tradicionalmente mais difíceis de prever.

Palavras-chave: Football match prediction, English Premier League, Machine Learning, Linear Models, Kernel Methods, Tree-based Methods, Neural Networks, Sentiment Analysis, Ensemble Methods, Hybrid Models

Acknowledgement

I want to thank my parents from the bottom of my heart for always being by my side and supporting me every step of the way. I am equally thankful to my uncles and my cousin, who always supported me and followed my progress month after month, even from a distance.

A very special thanks goes to my brother, whose encouragement was invaluable and who generously provided me with the computer that made it possible to train the models and fully embrace this study.

I would also like to leave a word of sincere appreciation for my supervisor. While searching for a topic that would be both challenging and meaningful, I proposed this idea, even without being very familiar with the subject at the time, and he immediately supported it, consistently providing the guidance and advice I needed to bring this work to a successful conclusion.

Finally, to all those who, in different ways, offered support, motivation, and trust during this journey, I extend my deepest thanks.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence.
ANN	Artificial Neural Networks.
ANOVA	Analysis of Variance.
API	Application Programming Interface.
AU-ARC	Area Under the Accuracy-Rejection Curve.
AUC	Area Under the Curve.
BERT	Bidirectional Encoder Representations from Transformers.
CatBoost	Categorical Boosting.
CPU	Central Processing Unit.
CSV	Comma-Separated Values.
CV	Cross-validation.
DAGs	Directed Acyclic Graphs.
DSR	Design Science Research.
ECE	Expected Calibration Error.
EPL	English Premier League.
FIFA	Fédération Internationale de Football Association.
FTR	Full-Time Result.
GA	Genetic Algorithm.
GDPR	General Data Protection Regulation.
GPU	Graphics Processing Unit.
GRU	Gated Recurrent Unit.
GWO	Grey Wolf Optimizer.
HTML	HyperText Markup Language.
HTTPS	Hypertext Transfer Protocol Secure.
ICA	Imperialist Competitive Algorithm.
kNN	k-Nearest Neighbors.
KS	Kolmogorov–Smirnov.
LightGBM	Light Gradient-Boosting Machine.
LLM	Large Language Model.

LSTM	Long Short-Term Memory.
MAE	Mean Absolute Error.
MiniLM	Minimal Language Model.
ML	Machine Learning.
MLP	Multi-Layer Perceptron.
MPNet	Masked and Permuted Pre-training for Language Understanding.
NB	Naive Bayes.
NGBoost	Natural Gradient Boosting.
NLP	Natural Language Processing.
NLTK	Natural Language Toolkit.
OOF	Out-Of-Fold.
PARX	Poisson Autoregression with eXogenous covariates.
PPV	Positive Predictive Value.
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
PSO	Particle Swarm Optimization.
QUOROM	Quality of Reporting of Meta-analyses.
R^2	R-squared.
RBF	Radial Basis Function.
RF	Random Forest.
RMSE	Root Mean Square Error.
RNN	Recurrent Neural Network.
RoBERTa	Robustly Optimized BERT Approach.
ROC	Receiver Operating Characteristic.
RPS	Ranked Probability Score.
SBERT	Sentence-BERT.
SFS	Sequential Forward Selection.
SHAP	SHapley Additive exPlanations.
SLR	Systematic Literature Review.
SMOTE	Synthetic Minority Oversampling Technique.
SVM	Support Vector Machine.
TF-IDF	Term Frequency-Inverse Document Frequency.
TNR	True Negative Rate.
TPR	True Positive Rate.
URL	Uniform Resource Locator.
VADER	Valence Aware Dictionary for sEntiment Reasoning.

WS	Window Size.
XAI	Explainable AI.
xG	Expected Goals.
XGBoost	Extreme Gradient Boosting.

Chapter 1

Introduction

This chapter provides an overview of the context for this study, establishing the relevance of the proposed hybrid predictive model. It will describe the focus of the research, present the central research question, and introduce the methodology that will guide the development and evaluation of the model.

1.1 Contextualization

Football, also referred to as soccer in some regions, is celebrated as the world's most popular sport, captivating billions of people around the globe. It is also a significant cultural and economic power with over 250 million players and 1.3 billion fans worldwide according to Fédération Internationale de Football Association (FIFA), having its impact not just confined to the playing field (Constantinou 2019). Among its many leagues, the English Premier League (EPL) is the biggest one as it is broadcasted to 643 million homes in 212 territories and has the capacity to reach a potential television audience of 4.7 billion people (Baboota and Kaur 2019). In the European football industry, the football market as a whole was projected to exceed 25 billion euros, and it is also part of the global sports betting market that has been estimated to be between 700 billion and $1 trillion dollars (Baboota and Kaur 2019; Constantinou 2019).

The use of statistical methods like Poisson regression and logistic models has been popular for modeling football match outcomes. However, these methods are limited in their ability to account for changing circumstances such as team morale and tactical changes (Beal et al. 2020).

New trends in the field of Machine Learning (ML) and Artificial Intelligence (AI) can also be useful in order to improve the accuracy of the predictions in the field of sports analytics. It is possible to enhance the prediction accuracy by integrating both the structured data and the unstructured data (e.g., sentiment from social media platforms) in a hybrid model to address the limitations of the traditional methods (Miranda-Peña et al. 2021).

Sentiment analysis has been found to be useful in capturing the emotional and psychological aspects that might affect the outcome of a match. For example, sentiment from social media reflects the mood and attitude of the fans, which may not be considered in a purely statistical analysis (Beal et al. 2020). These findings show that unstructured data can provide useful information that can be used in conjunction with the results of the traditional statistical analysis to better understand the factors that affect football matches (Kinalioğlu and Kuş 2023).

The EPL with its abundant data sources and global popularity, provides an ideal context to investigate the potential of hybrid models. However, the integration of structured and unstructured data in predictive analytics poses a number of methodological and practical challenges. These complexities must be addressed in order to develop more accurate and robust predictive models, which this study seeks to achieve by leveraging advances in machine learning and sentiment analysis.

1.2 Problem Statement

The prediction of football match outcomes remains a challenging task due to the sport's dynamic and unpredictable nature. Furthermore, the integration of structured data, such as match statistics and unstructured data including social media sentiment in football analytics is accompanied by several challenges owing to methodological and technical obstacles. Most approaches are not very efficient in merging different datasets, handling noisy or imbalanced sentiment data, and most importantly, they do not contain provisions for dynamic changes that may occur in real life such as changes in the team that start the game, or changes in the management of the teams involved in the game (Capobianco et al. 2019; Dip, N. Rahman, and Ahmed 2024; Miranda-Peña et al. 2021).

One of the key challenges in working with unstructured sentiment data is preprocessing, which is also constrained by imbalances, noise and irrelevant features, making the models developed to be unreliable and unscalable. Selecting important variables from both the structured and unstructured data requires a sophisticated feature selection technique, however, these techniques are very computationally intensive and are not scalable (Dip, N. Rahman, and Ahmed 2024; Miranda-Peña et al. 2021). This further increases the complexity of real time data processing as the predictive models have to be modified instantly in response to a change in the game such as a player getting injured or a change in the team or the manager (Capobianco et al. 2019; Hu and Fu 2022).

In addition, the lack of availability of large datasets that contain both structured and unstructured data also restricts the development of the hybrid model, its scalability and the generalization of models across different leagues and seasons (Dip, N. Rahman, and Ahmed 2024; Miranda-Peña et al. 2021). The following challenges are therefore crucial to address not only for improving the predictive performance in the context of the EPL and other highly competitive leagues, but also to advance methodologies in sports analytics.

1.3 Objectives

The primary objective of this study is to evaluate the efficiency of hybrid models that combine structured match data and unstructured sentiment data in predicting football outcomes, with a particular focus on the EPL.

To achieve this, the research seeks to answer the following question:

How effective are hybrid models in integrating statistical match data and sentiment analysis for football match outcome predictions compared to traditional methods?

In doing so, this study aims to provide fans, investors, and sports enthusiasts with a reliable tool for making informed decisions. By enhancing the credibility and robustness of football predictions, these hybrid models can also support team managers, analysts, and decision-makers in various strategic contexts.

To address the research question and achieve the main goal of this study, the following specific objectives have been defined:

- Conduct a comprehensive systematic literature review on hybrid predictive models applied to football match outcome predictions.
- Identify and collect relevant datasets, including statistical match data, and sentiment-related variables derived from social media and sports news, specifically for the EPL.
- Analyze existing machine learning techniques used in football outcome prediction, focusing on their limitations and the specific scenarios where they demonstrate advantages in hybrid modeling.
- Implement sentiment analysis techniques on social media posts and sports news to classify sentiments as positive, negative, or neutral.
- Integrate sentiment-based variables with traditional match data in a unified predictive model.
- Test the accuracy of hybrid models compared to traditional models based solely on statistical data.
- Evaluate the impact of sentiment analysis on the overall performance of hybrid models in predicting match outcomes.
- Explore the relative importance of different data types (statistical, sentiment) in the prediction process by employing machine learning techniques such as feature importance analysis and interpretability methods.
- Validate the proposed hybrid model using historical data from EPL matches.
- Propose actionable recommendations based on the findings to guide future research and practical applications of hybrid models in sports analytics.

1.4 Research Methodology

This study adopts the Design Science Research (DSR) methodology, which is particularly well-suited for addressing the objectives of this research. DSR is a research paradigm which is concerned with the development and evaluation of artifacts to address a particular problem in the world (Peffers et al. 2007).

The DSR methodology is most appropriate for this study for three reasons. First, DSR is concerned with the development of artifacts that are intended to be useful and innovative in addressing a given problem. In this research, the artifact is a hybrid predictive model that combines structured match data with unstructured social media sentiment to improve football outcome predictions.

Also, the development of the hybrid model requires the artifacts to be designed, the statistical and sentiment data to be integrated, and the model to be improved based on the experimental results, keeping in mind that DSR is an iterative process that improves the artifact.

Another key strength of DSR is its ability to bridge the gap between theory and practice. Moreover, DSR has a more rigorous evaluation process than other methodologies which is in line with the need to analyze the hybrid model in a systematic manner. Also, the methodology addresses the research problem, by providing a structured framework for designing, developing, and validating a hybrid solution, with all these steps aligned within the DSR framework.

Using the strengths of DSR in creating artifacts, improving through iteration, connecting theory and practice, and evaluating formally, this study offers a systematic method for developing and validating the hybrid predictive model.

The DSR methodology is composed of six iterative steps, which Peffers et al. (2007) describe. Each step is applied as follows:

1. **Problem Identification and Motivation:** The central problem addressed is the limitation of classical methods for predicting football match outcomes, which rely only on quantitative data and fail to consider qualitative information, such as social media sentiment. This limitation reduces prediction accuracy, which is not quite beneficial for stakeholders like team managers, analysts, and the sports betting companies.

2. **Defining Solution Objectives:** The main goal of this study is to create a hybrid predictive model that improves the accuracy of football match outcome predictions and combines structured match data and unstructured sentiment data, and to validate it in the context of the EPL.

3. **Design and Development:** The hybrid model will be built using the techniques of ML and Natural Language Processing (NLP). The focus will be on how to effectively combine the structured and unstructured data using the most recent methods.

4. **Demonstration:** The artifact will be demonstrated using real-world EPL match data combined with social media sentiment. Simulated scenarios will be employed to establish the model's capability of handling different types of data and applying them in real life situations.

5. **Evaluation:** The model's performance will be assessed based on certain criteria using quantitative measures. A comparative analysis will be made to establish how it performs in relation to other traditional models.

6. **Communication:** The conclusions and contributions of this research will be presented in this dissertation and other academic papers. The study is aimed at adding to the machine learning and sports analytics fields through the application of hybrid modeling strategies.

Chapter 2

State of Art

2.1 Introduction

Chapter 2 provides a wide range of information about the ideas, problems, and methods related to the research in question. The chapter consists of four main sections, which are designed to address essential elements of the study.

The first section is devoted to Machine Learning, focusing on the effectiveness and drawbacks of models for predicting football match results, the role of feature engineering, and the metrics used for model assessment. Next, the second section digs into Sentiment Analysis and how it can be used for football predictions, the difficulties of incorporating sentiment data, and the importance of feature engineering and evaluation metrics.

The third section is dedicated to Hybrid Models, covering their approaches, difficulties, evaluation, and specific application to sports forecasting. Last, the fourth section explains the Systematic Review, detailing the method, research questions, data sources, included and excluded studies, and the results of the review.

This chapter provides the necessary information for a conceptual and theoretical understanding of the hybrid models for football match outcome prediction developed in this research.

2.2 Machine Learning

Machine Learning is a subfield of artificial intelligence which deals with the development of algorithms that enable computers to learn from data and make predictions or decisions. It includes supervised learning in which models are trained on labeled data to make predictions about outcomes and unsupervised learning which is used to identify patterns in data without knowing what these patterns are (Nivetha et al. 2022). These techniques have been well applied in predicting football match outcomes using historical match statistics and other contextual data to enhance the accuracy of the predictions.

2.2.1 Models and Performance

It has been seen that various ML models have been applied for the purpose of predicting football outcomes with different levels of efficiency. The Random Forest algorithm is very popular and is often cited as one of the best. For instance, it produced a prediction accuracy of 68.55% in a study that analyzed ten seasons of EPL information, which was better than the other models, including C5.0 and Extreme Gradient Boosting (Raju et al. 2020). However, in another study it was found that Random Forest predicted only 68.9% accuracy for Type 1 matches which are of the kind where home team is likely to win (Ren and Susnjak 2022). Another paper also found that the Logistic Regression model provided satisfactory results since it was possible to accomplish 69.5% accuracy when predicting EPL using multi-season data (Elmiligi and Saad 2022). In situations with fewer features, CatBoost algorithm, a type of gradient boosting, has been found to be effective and it was able to predict 70% of Type 1 matches (Ren and Susnjak 2022). Other models like Support Vector Machine (SVM) and Gradient Boosting have been found to be less accurate with accuracy of 59% and 58.5% respectively (Raju et al. 2020). Overall, it has been observed that Random Forest and Logistic Regression models are frequently the best performers, especially when combined with proper feature selection and data cleaning.

2.2.2 Challenges in Football Match Outcome Predictions

Due to the complexity and unpredictability of the sport, several challenges are encountered in predicting football outcomes using ML. It is therefore difficult to model all the relevant variables accurately given that football matches are dynamic and are dependent on factors such as team form, injuries and tactical changes (Dip, N. Rahman, and Ahmed 2024; Elmiligi and Saad 2022). The problem of feature selection is another challenge since it is very important to identify the most relevant features from large datasets to improve the accuracy of the prediction. For instance, Analysis of Variance (ANOVA) F-tests and Random Forest based feature selection have been employed but the process is very complex and biased if not handled properly (Dip, N. Rahman, and Ahmed 2024; Elmiligi and Saad 2022). In addition, the efficiency of model performance is greatly determined by the quality and completeness of the datasets used. Studies have been criticized for using small or biased datasets that fail to consider important elements such as player performance measures or other aspects like weather and venue (Dip, N. Rahman, and Ahmed 2024; Rose et al. 2022). Model performance is also variable across leagues and seasons, hence, the models may not be able to generalize to data collected at a different time or in a different league than the one they were trained on (Da Costa, Prudêncio, and Mota 2023). In addition, the fact that football teams are dynamic (changes in form, fitness, or even the manager or the team) calls for constant model reformulation (Rose et al. 2022).

2.2.3 Role of Feature Engineering

Feature engineering is a crucial process that enhances the performance of ML models for football match predictions. This transforms raw data into improved inputs that increase model accuracy and interpretability. In football, this means creating new features including team form, league points and player statistics, all of which are better suited to the sport than raw data (Dip, N. Rahman, and Ahmed 2024; Malamatinos, Vrochidou, and Papakostas 2022). For instance, an investigation on the Greek Super League revealed that engineered features like Home Team Form and Away Team Form do the job efficiently and increase the accuracy of the predictions (Malamatinos, Vrochidou, and Papakostas 2022). For instance, feature selection using ANOVA in the Italian Serie A League identified 28 out of 54 features as effective, which improves the likelihood of classification (Taşpinar, Çinar, and Koklu 2021). Other advanced techniques like Sequential Forward Selection (SFS) and data augmentation methods like Synthetic Minority Oversampling Technique (SMOTE) have also been used in an effort to reduce the effects of class imbalance and overfitting and thus improve model robustness (Malamatinos, Vrochidou, and Papakostas 2022). Feature engineering enables models to capture important aspects like home advantage and team dynamics because it is incorporated with domain knowledge to make accurate predictions (Taşpinar, Çinar, and Koklu 2021).

2.2.4 Evaluation Metrics

To guarantee both accuracy and reliability, robust metrics are required for evaluating ML models in football predictions. Accuracy is the most often used metric, which presents the proportion of correct predictions. To this end, it is important to note that for imbalanced datasets, measures such as precision, recall, and F1 score are frequently used (Dip, N. Rahman, and Ahmed 2024; Ren and Susnjak 2022). Particularly, when false positives and false negatives have different consequences, the F1 score that is between precision and recall is especially useful to evaluate the models (Dip, N. Rahman, and Ahmed 2024). Another widespread metric is the Area Under the Curve-Receiver Operating Characteristic (AUC-ROC) (AUC), which indicates how effectively a model can separate between classes (Baratela et al. 2024). Some studies have also applied the Area Under the Accuracy-Rejection Curve (AU-ARC) to evaluate the models with a rejection option to improve the reliability of predictions by excluding uncertain predictions (Da Costa, Prudêncio, and Mota 2023). Sometimes, absolute and squared errors, such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), are used to assess the models that produce continuous outcomes, e.g., match scores (Dip, N. Rahman, and Ahmed 2024). These metrics allow for a thorough assessment of model performance, which guarantees that the ML models for football predictions are accurate, robust, and reliable in different situations and datasets.

2.3 Sentiment Analysis

Sentiment Analysis is a computational technique used to determine the polarity of text, i.e., to categorize it as positive, negative, or neutral. It is developed to extract the subjective information from the textual data to understand the public opinion or emotional tone. In the context of football predictions, the sentiment analysis is applied on social media platforms like X (formerly known as Twitter) to understand what people feel about a match or a player. Techniques like tokenization and lemmatization (performed by tools like Stanza) are used for sentiment classification, where prediction polarity dictionaries have been applied to associate positive words with positive outcomes and negative terms with negative sentiment (Miranda-Peña et al. 2021).

2.3.1 Approaches for Football Predictions

The sentiment analysis has been used in football outcome predictions based on the sentiment of fans and experts from social media and media articles. Various studies have employed machine learning methods, including SVM and Random Forests, to categorize the sentiments and incorporate them in the predictive models. For instance, one study has used tweets and media articles to include psychological factors like rivalries and fan mood in the machine learning models and found that they enhance the predictive accuracy (Beal et al. 2020; Miranda-Peña et al. 2021). These approaches illustrate how sentiment analysis can be combined with statistical methods to enhance predictions, and some studies have reported that this integration can lead to as high as 30% profitability for EPL matches (Da Costa, Prudêncio, and Mota 2023).

2.3.2 Challenges in Football Predictions

Although promising, there are some issues with applying sentiment analysis to football predictions. The nature of the sport being extremely unpredictable, combined with the fact that fan sentiments are essentially subjective, results in noisiness and imbalanced datasets (Miranda-Peña et al. 2021). Furthermore, combining sentiment data with structured match statistics is an computationally expensive task that requires sophisticated methods of preprocessing of sentiment data (Tammouch, Elouafi, and Essadik 2024). This makes it complicated to rely on sentiment analysis alone because of the dynamic nature of football, which may be transformed by factors like player injuries or tactical changes (Da Costa, Prudêncio, and Mota 2023).

2.3.3 Feature Engineering

Feature engineering is the process of transforming raw sentiment data into useful variables that improve the accuracy and robustness of predictive models. In football predictions, this involves developing measures like player sentiment scores or fan mood. Methods like SFS and other data augmentation techniques, such as SMOTE, have been used to improve model performance through dealing with the class imbalance problem (Malamatinos, Vrochidou, and Papakostas 2022; Verdonck et al. 2024). This allows the models to understand the complicated sentiment trends in terms of home advantage or team dynamics, thus leading to better predictions (Taşpinar, Çinar, and Koklu 2021).

2.3.4 Evaluation Metrics

The efficiency of the sentiment analysis models has to be assessed using a proper set of measures to ensure their effectiveness in the context of football predictions. Although accuracy is still widely used, it is usually complemented by precision, recall, and F1-score to account for the imbalanced sentiment distributions (Dip, N. Rahman, and Ahmed 2024; Ren and Susnjak 2022). The Bidirectional Encoder Representations from Transformers (BERT) and the Robustly Optimized BERT Approach (RoBERTa) frameworks, which are advanced frameworks, work well but require a lot of computational power, which makes it possible to discuss the tradeoff between efficiency and accuracy (F. Wang 2024). Confusion matrices also assist in identifying regions for improvement and provide a more detailed view of model performance across different sentiment classes (Iyer 2024).

2.4 Hybrid Models

Hybrid models in machine learning are new and efficient methods that are developed from combining different algorithms or techniques in order to improve the predictive capability by benefiting from the strengths of individual models while mitigating their limitations. These models are very useful for complex prediction problems in different areas. For example, it has been established that combining SVM and Naive Bayes (NB) is better than using a single algorithm in a classification task. A study comparing SVM-NB hybrids and Lasso-Ridge techniques shows that the former achieve higher predictive accuracy in real-world applications (Abdullah et al. 2024).

2.4.1 Hybrid Models Approaches

Hybrid models in machine learning are designed to take advantage of multiple techniques, such as using optimization techniques and integrating model-based and data-driven techniques. One of the most popular techniques is combining several machine learning techniques to take advantage of their individual qualities. For example, it has been shown that combining SVM with decision trees is better than using a single algorithm in terms of predictive accuracy (Byeon 2021). Other methods like stacking and boosting improve the robustness of the model and are particularly important when dealing with imbalanced datasets (T. Chen et al. 2019; R. Rahman et al. 2024). A combination of several models produces a better classification accuracy and is more generalizable than a single model (Amir Mosavi et al. 2019). A critical part of the development of hybrid models is the incorporation of optimization techniques. The Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) are commonly used to tune the parameters of the model and improve both accuracy and speed (Madeswaran et al. 2023; Amir Mosavi et al. 2019). These optimization techniques contribute to model interpretability, which is also important in practice, especially in healthcare and finance (Giroh, Kumar, and G. Singh 2023). Hybrid models are also applied to combine the data-driven and model-based strategies. In the energy systems, the performance of physical models combined with machine learning techniques has been better because these methods are able to capture both the empirical data and theoretical knowledge (Amir Mosavi et al. 2019). This integration is most useful in real-world applications that require a mix of practical data and theoretical information. In addition, the development in computational methods has made it possible to combine deep learning with conventional machine learning techniques, thereby advancing the field of hybrid models. For instance, hybrid quantum neural networks have been successfully applied to high-dimensional data classification, which points to the possibility of applying novel hybrid models (H.-Y. Chen et al. 2023). In summary, hybrid models are a collection of methods that combine different techniques to tackle the problems presented by the data, using an optimization technique to adjust the parameters of the model to enhance its performance. The combination of these elements provides a powerful framework for a wide range of applications, including sports analytics, healthcare, industrial processes, ultimately improving predictive performance.

2.4.2 Hybrid Models Challenges

The application of hybrid models presents several challenges, both methodological and technical, as discussed below. These challenges are due to the combination of different techniques, interpretability, data collection, and optimization issues. A major challenge is combining several techniques from machine learning to achieve their best results. Although hybrid models are usually better than single approach-based models, they can be very difficult to design and implement. For example, most hybrid models are based on a single ontological framework, and this can become unwieldy and difficult to manage, especially in a multi-domain context. Furthermore, using a single algorithm on different data sets can be not very efficient and therefore, the choice of the algorithm and its parameters should be studied in order to achieve the desired effectiveness across all environments (Narsis, Dujardin, and Nicolle 2023). Another key problem is the interpretability of the hybrid models. Many such models are black boxes, not providing clear information on how their decisions are made. This lack of transparency is particularly problematized in practice within applications like healthcare and finance where decision making has real life implications. Explainable AI (XAI) has become an important feature such that, besides accuracy, hybrid models should also provide a clearness on how predictions are made (Giroh, Kumar, and G. Singh 2023).

Data management also poses a problem for the development of hybrid models. Preprocessing and feature selection for large and heterogeneous datasets can be time consuming and computationally expensive. As Dahiya, Handa, and N. P. Singh (2015) pointed out, proper input data is crucial as it directly affects the model's performance, and biased training should be avoided. In addition, combining data from different sources increases computational requirements, and, therefore, the concern about the scalability and efficiency of hybrid models arises (Narsis, Dujardin, and Nicolle 2023). Optimization techniques, including metaheuristic algorithms, have been applied to enhance the performance of hybrid models, but their application presents another level of complexity. Such algorithms are typically computationally expensive and require adjustment to ensure that the hybrid framework works well (He et al. 2023). For example, designing a hybrid model with several levels of algorithms requires very careful planning and exhaustive testing, which can be time and resource consuming (Pang 2023). In summary, although hybrid models are very promising, they are also difficult to develop and apply. These difficulties are connected with the combination of different methods, techniques, data sets, and optimization problems. These challenges must be overcome to successfully use hybrid models in practical applications and ensure their reliability and credibility across various fields of study.

2.4.3 Hybrid Models Evaluation

Hybrid models of machine learning are assessed for their performance and potential use in real world applications in this study. These models are designed by combining several algorithms and methods, and they are evaluated using several criteria and evaluation procedures to determine their predictive properties. RMSE, MAE, and R-squared (R^2) are some of the most frequently used metrics to assess the goodness of fit of the models, which give information about the quality, reliability and explanatory power of these models. R-squared is a popular metric that is used to establish the goodness of fit of a model that is a measure of the proportion of variance in the dependent variable that is explained by the independent variables. This metric is particularly useful for hybrid models used in complex systems like financial forecasting and energy systems to understand the model's explanatory power (Amir Mosavi et al. 2019).The RMSE is one of the most frequently used metrics and is a measure of the average magnitude of error between predicted and actual values, with smaller values being better. Nosratabadi, Amirhosein Mosavi, et al. (2020) showed that hybrid deep learning models are better than single deep learning models in terms of RMSE, which means that they are more accurate in their predictions. Moreover, the use of RMSE in conjunction with other measures such as MAE offers a more accurate assessment. The absolute error, independent of its sign, is measured by MAE and is particularly useful in applications where overestimation and underestimation are penalized equally. For example, MAE has been used to evaluate the prediction errors of hybrid models, such as Artificial Neural Networks (ANN)-Imperialist Competitive Algorithm (ICA) and ANN-Grey Wolf Optimizer (GWO), for crop yield prediction (Nosratabadi, Szell, et al. 2020). Besides these metrics, cross validation is widely used to ensure that hybrid models are not overfitted to the training data and can generalize well to new data. Cross validation is a technique of dividing the dataset into several subsets, training the model on some of them and checking it on others, which helps in preventing overfitting and guarantees the stability of the model. This approach is most useful for hybrid models that combine different algorithms because it allows the researchers to determine which component is vital to the model's effectiveness (Pang 2023). Ensemble techniques that are a form of hybrid modeling are also under consideration. Methods like stacking and boosting work by combining the predictions from multiple base models. For example, ensemble techniques have been found to be better than individual models as they are able to leverage the strength of each model and at the same time, weaken its weakness thus improving accuracy and reliability (Gorczyca, Toscano, and Cheng 2019). Studies have applied ensemble strategies to hybrid models and found that they are effective for combining predictions and increasing model stability (Kazienko, Lughofer, and Trawinski 2015).

In conclusion, the assessment of hybrid models is a multifaceted problem that uses techniques such as RMSE, MAE and R^2 along with other methods like cross validation and ensembling. These methods provide a complete framework for evaluating the performance of hybrid models, fine-tuning them, and using them in practice.

2.4.4 Hybrid Models in Sports Prediction

Hybrid models have gained much attention in sports prediction, especially for the purpose of predicting football match outcomes since they combine traditional statistical methods with various technical elements of machine learning. These models combine various data sources and methods, therefore increasing the accuracy of the prediction and solving the problems of traditional methods. The Poisson Autoregression with eXogenous covariates (PARX) model is a good example, which includes exogenous variables in a Poisson autoregressive framework to model the distribution of goals. This method captures from the attacking and defensive side, form and other factors and thus gives a better picture of the game than models that are limited to historical data only (Angelini and Angelis 2016). Similarly, the Dolores model uses Hybrid Bayesian Networks and dynamic ratings to enhance prediction accuracy across leagues, and it shows how combining cross-league data improves the generalization capability (Constantinou 2019). Gated Recurrent Units (GRU) that are advanced forms of Recurrent Neural Network (RNN) have also been successfully used in determining the winner of a match by analyzing player and team ratings (AlMulla et al. 2023). These models are however better at capturing the complex patterns such as the team dynamics and the player performance that are difficult to capture by traditional methods (Peters and Pacheco 2023). Therefore, hybrid models are further developed to focus on feature engineering and complex network metrics. For example, combining gradient boosting classifiers with neural networks has been useful in predicting professional American football games player performance and finding that these models can be better than baseline models (Nimma and Uddagiri 2024). For instance, integrating passing network metrics with traditional match data leads to better predictions and provides additional insight into the tactics of the teams and movements of players (Baratela et al. 2024). These methods also show that hybrid models are adaptable and can incorporate different features for improved predictive capabilities. This adaptability of hybrid models is particularly important in a dynamic environment, such as sports. Thus, hybrid approaches are better in dealing with problems like overfitting and predicting a draw than traditional models like Random Forest and SVM (T. Wang, Zhang, and Zhu 2024). Also, hybrid models allow the analysis of match outcomes in a more general manner, not only the prediction of the match result, but also the tactics for the match. For example, the TacticAI model that combines predictive analytics with generative modeling to show how hybrid approaches can assist football teams with their strategic planning (Z. Wang et al. 2023). In addition, it has been shown that Large Language Models (LLM) can be as effective as traditional machine learning methods, such as Random Forest and Extreme Gradient Boosting (XGBoost), in football prediction, and with the advantage of lower computational costs and simpler implementation since LLMs do not require model training (J. Li, Zhao, and Z. Li 2024). In conclusion, hybrid models are an enhancement of traditional statistical methods combined with advanced machine learning techniques for the purpose of sports prediction. Since they are able to work with various data sources, are adaptable to changing environments, and are able to provide additional information about the game, they become a useful tool for improving the accuracy of predictions and decision making in the field of sports analytics. These advancements not only increase the accuracy of predictions but also widen the scope of analysis, making hybrid models valuable in shaping and understanding sports outcomes.

2.5 Systematic Review

Systematic Literature Review (SLR) is a rigorous and structured process of identifying, evaluating and synthesizing research in order to address a particular question, topic or subject in the literature. Not like conventional reviews, an SLR has a clear set of guidelines that any person can follow and which make the review process unbiased and transparent. Through the assessment of primary studies, it creates a secondary study that offers an exhaustive view of the research in question.

The main goal of this study is to present an overview of the findings that have been obtained in the course of the research, and to assess the strengths and weaknesses of the methods used. Additionally, it reveals the gaps in the research that point to where future studies should be directed. SLRs are also important in ensuring that new research is introduced within the context of previous research so as to avoid repeating what has already been done. In addition, SLRs are a valuable tool for testing theoretical constructs with empirical data and developing new ideas.

The present systematic review strives to capture all available literature related to the research questions and integrate them to advance the knowledge in this specific area of study.

2.5.1 Methodology

For this study, the systematic literature review will follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology with all its components, which is a strong framework to enhance the transparency, clarity and completeness of systematic reviews and meta-analyses (Moher et al. 2009). PRISMA is made up of a 27-item checklist and a four-phase flow diagram that assists researchers in identifying the different stages of the review in order to enhance reporting. Initially developed from the Quality of Reporting of Meta-analyses (QUOROM) checklist in 2005, PRISMA emphasizes iterative processes and adaptability to various research contexts, making it suitable for this work on hybrid models in football match predictions.

This study will do the following for instance, based on PRISMA guidelines: develop research questions that are central to the objectives, conduct a comprehensive search of the databases and input search terms, apply inclusion and exclusion criteria to the selected studies, assess the quality of the selected studies using predefined assessment criteria, and extract, synthesize, and present findings. This is a systematic approach to ensure that reliable and comprehensive insights into the literature are derived.

However, it's worth noting that while this approach provides a systematic and comprehensive framework for the literature review, it may also involve challenges such as managing a large volume of studies, handling inconsistencies in study design and methodology, and synthesizing findings effectively. By carefully following the PRISMA guidelines and addressing these challenges proactively, this study aims to provide valuable insights into the current state of research on hybrid models in football match predictions and contribute meaningfully to the field of study.

2.5.2 Research Questions

The primary goal of this research is to develop and assess hybrid models that combine the use of structured and unstructured data to forecast football match outcomes. This research responds to the methodological, technical, and evaluation challenges of hybrid approaches with the ultimate goal of contributing to the field of sports analytics and increasing predictive accuracy. From this objective, the following research questions have been formulated:

RQ1: What hybrid methods have been utilized to combine structured and unstructured data for predicting football match outcomes?

RQ2: What are the primary methodological challenges in integrating structured and unstructured data in predictive models for football match outcomes?

RQ3: What metrics are commonly used to evaluate the effectiveness of hybrid models in predicting football match outcomes?

RQ4: What are the limitations and opportunities in using hybrid models for predicting football match outcomes?

2.5.3 Data Sources

For this systematic literature review, the selected data sources include Academic Search Complete[1], IEEE Xplore[2], Science Direct[3] and Web of Science[4]. These databases were chosen because they cover thoroughly peer reviewed articles in areas very closely related to this research, such as machine learning, sentiment analysis, and hybrid modeling. IEEE Xplore has high quality resources on computational methods and advanced technologies, Science Direct offer interdisciplinary information with respect to sports analytics. Web of Science allows access to influential and highly cited studies and Academic Search Complete offers a wide range of academic articles across all disciplines. The above mentioned databases are rich in high quality research materials and hence are suitable for conducting this study in a proper manner.

2.5.4 Search Terms

The purpose of selecting search terms for this systematic review was to guarantee that all available literature that is most relevant to the research questions would be retrieved, without any unnecessary bias. The keywords which were chosen were meant to encompass the main categories of the research: football match outcome predictions, hybrid models and the inclusion of both structured and unstructured data. The context was covered by the terms "football", "soccer", and "Premier League", while the predictive focus of the study was captured by "predict*", "forecast*", and "outcome". The study also employed advanced methodologies such as machine learning, artificial intelligence, AI, sentiment analysis and deep learning hence the use of keywords like these. To narrow or expand the search,

[1] https://www.ebsco.com/products/researchdatabases/academic-search-complete
[2] https://ieeexplore.ieee.org
[3] https://www.sciencedirect.com
[4] https://www.webofscience.com

Boolean operators (AND, OR) were used in order to achieve the desired level of specificity or inclusivity in the search. The use of truncation (for example, "predict*") was to ensure that all possible derivatives of the search terms were captured, such as "prediction" and "predictive". Moreover, the query syntax was developed and fine-tuned during the course of preliminary searches to increase the likelihood that the retrieved articles are relevant. For instance, adjustments were made to align the query with database-specific requirements, as some databases have limits on the number of Boolean operators or wildcards allowed.

The search query utilized for this study is as follows:

("football" OR "soccer" OR "Premier League") AND ("predic*" OR "forecast*" OR "outcome") AND ("machine learning" OR "artificial intelligence" OR "AI" OR "sentiment analysis" OR "deep learning")

Table 2.1: Search terms

Scope	**String**
Football	("football" OR "soccer" OR "Premier League")
Match Prediction	("predic*" OR "forecast*" OR "outcome")
Artificial Intelligence	("machine learning" OR "artificial intelligence" OR "AI" OR "sentiment analysis" OR "deep learning")

The major limitation of the study was that many irrelevant studies were obtained from the broad terms "football" and "forecast". These problems were solved by developing better search terms and adding more specific keywords such as "machine learning" to limit the search. In conclusion, the search terms were properly identified to directly address the research questions with due consideration to the conventions of the databases used. This approach led to the identification of a large number of high-quality studies to be included in the systematic review.

2.5.5 Inclusion and Exclusion Criteria

The following inclusion criteria were applied to ensure the relevance and quality of studies for this systematic review:

- The source explicitly addresses football (soccer) match outcome prediction (win, draw, or loss), especially in major leagues (e.g., Premier League, top-tier European leagues).
- The source demonstrates how match statistics (e.g., historical performance) and/or sentiment data (e.g., social media, sports news sentiment) are utilized or combined for predictive modeling.
- The source presents quantitative performance results (e.g., accuracy, precision, recall, F1-score) or other validation metrics.
- The source compares its approach with traditional methods (purely statistical) or discusses the limitations of single-source data methods.

The following exclusion criteria were applied to ensure the selected studies meet the objectives and scope of this review:

- Sources not written in English.

- Sources published before 2019.
- Sources besides journal articles, chapters, conference proceedings, and books.
- The source does not utilize machine learning, sentiment analysis, or a hybrid approach (e.g., purely qualitative analyses, descriptive statistics without predictive modeling).
- The source does not focus on football (soccer) or deals purely with fan behavior, marketing, or economics without addressing match outcome prediction.
- The source fails to describe its data sources or predictive modeling approach with sufficient clarity.

To ensure consistency in interpretation and evaluation, only English language sources were included as a translation may introduce inaccuracies or misinterpretations. The decision to exclude sources published before 2019 is because of the rapid advancement in machine learning and hybrid models in recent years, such that only studies that incorporate the last methodology and technology are considered. This cutoff ensures the review remains relevant to current practices and state-of-the-art techniques.

2.5.6 Quality Assessment

After the papers passed the inclusion and exclusion criteria, their quality was assessed based on the following aspects:

- Each study was evaluated for its contribution to the diversification of effective prediction techniques.
- The studies were analyzed to ensure they present well-defined methodologies, detailed performance metrics, and clear conclusions.
- The citation count of each paper was reviewed to provide insight into its impact and recognition within the academic community.

This quality assessment ensured that the selected studies contributed meaningful insights and methodological rigor to advance prediction techniques in the context of football match outcomes.

2.5.7 Data Extraction and Synthesis

A total of 697 records were retrieved from the initial search from four databases: Academic Search Complete (n=170), IEEE Xplore (n=221), Science Direct (n=27), and Web of Science (n=277). First, the duplicate entries were removed and 607 records were found to progress to the screening phase. From the 509 reports excluded during the screening based on the title and abstract, 98 were taken forward for detailed eligibility assessment. The eligibility assessment involved the application of predefined inclusion and exclusion criteria (IC1-IC4, EC1-EC6), which led to the exclusion of additional 45 reports. From this phase, 62 studies were considered eligible for inclusion. Also, 9 records were found to be potentially relevant from other websites and were incorporated into the list of included studies to ensure that all relevant studies were included. Figure 2.1 demonstrates the systematic selection process from identification to inclusion, which yielded a final dataset of 62 studies for this systematic review.

2.5.8 Research Questions' Answers

After the process of data extraction and synthesis, the papers were analyzed in detail in order to respond to the research questions. They were carefully reviewed in order to get a full picture of their content and for the following conclusions to be made.

RQ1: What hybrid methods have been utilized to combine structured and unstructured data for predicting football match outcomes?

Hybrid methods are a good way of combining the two types of data that are both structured and unstructured in order to predict the outcome of football matches. This combination is useful in that it captures the best aspects of both types of data in the quest to create a more holistic model of football match dynamics.

The most widely used approach is to combine machine learning techniques with statistical analysis to analyze data of different types. For instance, for structured data analysis, researchers have applied Random Forest, Support Vector Machines, and Gradient Boosting algorithms which have produced promising results (Dip, N. Rahman, and Ahmed 2024; Iyiola, Okagbue, and Odetunmibi 2022). Pedagogical recommender systems, including RNN and Long Short-Term Memory (LSTM), have been proposed to incorporate the sequential structured and unstructured data with the latter being used to model the temporal characteristics of football match outcomes (Nivetha et al. 2022; Sreenivasgoud et al. 2024).

This is because information such as fan's sentiment analysis from social media platforms provides a more real-life view of the match. In this paper, K-means and fuzzy C-means clustering algorithms in conjunction with ANNs and Random Forest (RF) have been used to analyze fan sentiments and match data to enhance the prediction model (Kinalioğlu and Kuş 2023).

Feature engineering is also very essential in hybrid methods. Some techniques have been used in feature selection such as ANOVA F-tests and some new data preprocessing techniques such as calculating the confidence rating of the historical data to combine both the structured and unstructured data in the modeling process (Dip, N. Rahman, and Ahmed 2024; Kinalioğlu and Kuş 2023).

Beyond these examples, research into hybrid methods has expanded to new directions through recent studies. Beal et al. (2020) used The Guardian pre-game previews together with structured match statistics to achieve a 6.9% better predictive accuracy than statistical models alone. Wunderlich and Memmert (2022) analyzed 1.9 million tweets together with in-play data yet discovered that social media failed to outperform betting odds in predictive accuracy. Yeung, Bunker, and Fujii (2023) created a system which united FIFA player ratings with coaches' tactical strategies to enhance both model readability and forecasting precision.

The implementation of hybrid approaches in modeling combines different analytical methods. Elmiligi and Saad (2022) developed a statistical–machine learning hybrid system which processed 200,000 matches to detect both statistical patterns and non-linear trends. Razali et al. (2022) combined pi-ratings with TabNet deep learning models to achieve better accuracy and RPS performance. Miranda-Peña et al. (2021) demonstrated that fan expertise features extracted from social media platforms enhance the predictive power of odds and statistical data in pre-match forecasting.

These hybrid methods are found to be more efficient than the conventional systems. For instance, studies that combined machine learning and sentiment analysis achieved better prediction accuracies and the accuracy increased with the different match outcomes (Kinalioğlu and Kuş 2023). Nevertheless, there are still many possibilities available in the area of research. For instance, the use of unstructured data such as news or fan interactions is an area that has not been explored to the fullest. In the future, it will be important to apply NLP techniques to the textual data to gain more meaningful information for the prediction process.

In conclusion, it can be said that combinations of statistical analysis, machine learning methods, and data management are successful in identifying football match results. These approaches enable the combination of various data sources to enhance the quality of the predictions and to comprehend the nature of the match.

RQ2: What are the primary methodological challenges in integrating structured and unstructured data in predictive models for football match outcomes?

The use of structured and unstructured data in predictive models for football match outcomes is quite challenging. These challenges result from the type of data available, the dynamics of the football game, and the need to advanced approaches in data integration and predictive modelling. One major challenge is data collection and integration. Typically, predictive models are developed using a variety of factors including historical match data, player performance data, and other conditions like weather and home ground advantage. Nevertheless, the combination of data across various leagues and years produces a wide range of data availability and accuracy which can lead to uncertain model results (Berrar, Lopes, and Dubitzky 2019). Furthermore, missing data is an issue that comes with combining different data sets and this missing data can lead to bias in the predictions if it is not dealt with properly. To address these gaps, data imputation and ensemble models are usually used; however, this increases the complexity of the modelling process (Dip, N. Rahman, and Ahmed 2024).

Another challenge is feature engineering because it involves the process of selecting relevant derivatives from the data. This is especially a challenge when trying to combine numerical data, which is readily available and easily codified, with unstructured text such as match reports. However, to reduce the number of features, ANOVA F-tests and feature selection based on the Random Forest algorithm may be required, but all these need to be validated to ensure the accuracy of the model (Dip, N. Rahman, and Ahmed 2024). In addition, the dynamics of the football makes it difficult to have constant and well-defined features that can be used in the prediction of the outcomes (Macrì Demartino, Egidi, and Torelli 2024).

The process of extracting predictive signals from unstructured data sources presents an additional challenge. The extensive amount of data available on social media platforms does not necessarily translate to useful predictive capabilities. Wunderlich and Memmert (2022) demonstrated that Premier League match tweets mirrored supporters' perspectives yet failed to enhance prediction accuracy beyond what structured pre-game betting odds provided. Moreover, the conversion of this sentiment into workable variables proves challenging because it needs sophisticated preprocessing which might lead to data contamination and biased results (Miranda-Peña et al. 2021).

Model development and validation are also not without some hurdles. To integrate the two types of data, complex tools may be required, including neural networks and ensemble methods, which are costly to compute and susceptible to overfitting (Rodrigues and Pinto

2022). The model generalizability and the appropriate evaluation metrics including accuracy, precision, and Ranked Probability Score (RPS) remain important but difficult to achieve (Constantinou 2019).

Beal et al. (2020) demonstrated that using match statistics with journalistic articles produced better predictions, yet using only one media source limited the ability to generalize results. Malamatinos, Vrochidou, and Papakostas (2022) highlighted that models trained on one league fail to perform well in other competitions because different leagues have unique characteristics.

In addition, the analysis of structured datasets reveals multiple restrictions in their data collection process. The predictive power of unstructured data becomes less reliable when using incomplete statistics and biased sample data that tends to show stronger teams (Dip, N. Rahman, and Ahmed 2024). The practical application of these models faces two main obstacles because they lack clear explanations and simple usage methods. Yeung, Bunker, and Fujii (2023) presented a method which combined FIFA player ratings with coaches' tactical plans to enhance performance, yet they emphasized that coaches and analysts need models that deliver accurate predictions while remaining easy to understand.

The main difficulty in multimodal fusion remains a significant challenge. The combination of different data types including statistics, odds, player information, and text becomes extremely challenging to integrate because of their diverse nature and unclear meanings (Zheng 2023). Elmiligi and Saad (2022) revealed that feature engineering stands as the main obstacle for developing stable hybrid models that handle extensive datasets.

In conclusion, these challenges underscore the importance of innovative data integration strategies, complex feature construction, and sophisticated algorithms. For this reason, overcoming these methodological barriers is crucial to improving the predictive power of hybrid models in football match outcome prediction.

RQ3: What metrics are commonly used to evaluate the effectiveness of hybrid models in predicting football match outcomes?

Analyzing the efficiency of the hybrid models in predicting football match outcomes is a quantitative exercise that requires the use of a number of accuracy and quality measures. These metrics help the researchers in assessing the performance of the models and, in turn, help in identifying the scope of enhancement.

Accuracy is a simple yet effective measure that tells us how accurate is the model in identifying the correct output from the given input. It is mainly used to assess the results of hybrid models that use both structured and unstructured data (Elmiligi and Saad 2022). For example, a Gaussian Naïve Bayes model reached 85.43% accuracy which exceeded the 79.81% performance of the Decision Tree classifier according to P, D, and S (2023). The evaluation of imbalanced datasets requires additional metrics because accuracy alone does not provide sufficient information.

Besides accuracy, the F1-score that is the harmonic mean of precision and recall gives a single value for both false positives and false negatives where there are imbalances in classes (Kinalioğlu and Kuş 2023). The study by Capobianco et al. (2019) achieved 0.857 precision and 0.750 recall when they applied Random Forest for winner prediction. The interpretable framework developed by Yeung, Bunker, and Fujii (2023) achieved an F1-score of 0.47 through the combination of FIFA ratings and tactical formations which outperformed betting odds at 0.39. Research studies use multiple indicators that extend beyond traditional measures. The research by Kinalioğlu and Kuş (2023) evaluated hybrid clustering-classification models through multiple performance metrics including True Positive Rate (TPR), True Negative Rate (TNR), Positive Predictive Value (PPV), F-measure, Kolmogorov–Smirnov (KS) and accuracy which resulted in maximum of 81.76%.

The RPS metric is based on the idea of measuring the distance between the predicted probabilities and the real outcomes. A lower RPS indicates that there is a better match between the predictions and the actual events. For instance, hybrid models that use CatBoost have been found to be accurate with low RPS values, which are useful for making probabilistic assessments of the matches in detail (Nazim Razali et al. 2022). The TabNet deep learning model with pi-ratings developed by Razali et al. (2022) achieved better accuracy and RPS results than previous ensemble methods according to their research. Similarly, Baboota and Kaur (2019) applied gradient boosting in the English Premier League and reported an RPS of 0.2156, benchmarking it directly against bookmaker odds.

Precision is the proportion of positive predictions that are actually positive, and it is a measure of how accurate a model is in identifying particular events, for instance, wins or losses. Recall is a complementary measure to precision that helps in determining the ability of the model to assign all the relevant instances to a particular class. Moreover, the AUC is a measure that helps to understand how well a model can separate between classes, giving a general idea of the quality of the predictions (Baratela et al. 2024). Malamatinos, Vrochidou, and Papakostas (2022) demonstrated that the "draw" class recall performance improved substantially following the optimization of hyperparameters. Ren and Susnjak (2022) highlighted that explainable hybrid models require evaluation through the combination of accuracy, precision, recall and F1-score according to the Kelly Index.

The confusion matrices give a clear picture of the true positives, false positives, true negatives and false negatives to check the accuracy of the model for many classes, e.g., home win, draw, and away win (Hassard and Kerr 2024). More sophisticated metrics, such as SHapley Additive exPlanations (SHAP) values and permutation importance, are gradually being adopted to explain the role of each feature in the model prediction. These methods improve the explainability and performance of the hybrid models through the feature importance interpretation (Baratela et al. 2024).

Simulation is a commonly used technique in hybrid models to determine the probabilities of various outcomes, for instance, tournament winners or particular match results. This approach gives a holistic view of the possible situations and confirms the validity of the probabilities assigned by the model (Groll et al. 2019). Additionally, some works also analyze class-specific performance. For example, convolutional neural networks reached an overall accuracy of 55.31%, but with 71.52% accuracy specifically for predicting home team wins (Satyapanich and Somkheawwan 2023).

All these metrics, from accuracy (the simplest) to SHAP values or permutation importance (the most intricate), provide a clear and coherent approach to assessing hybrid models. Therefore, using these metrics, the researchers can produce efficient and precise forecasts with many opportunities for model enhancement.

RQ4: What are the limitations and opportunities in using hybrid models for predicting football match outcomes?

Recent studies show both the limitations and potential of hybrid models for predicting football match outcomes. The first major limitation is that football matches are generally complex and unpredictable processes that are a function of a number of factors, including team strategies, player ability, injuries, and even conditions beyond the field. Because of the randomness of the processes, it is rather difficult to ensure that hybrid models are accurate in a wide variety of settings. For example, the problem of predicting draws is still present, and many hybrid models, including those that use sophisticated approaches, fail to provide accurate classification of this result (Gudla et al. 2023). Furthermore, hybrid models are generally more computational in nature and tend to combine different algorithms and large amounts of data, which can lead to issues with regard to scalability and real-time application, particularly when handling both structured and unstructured data (Rodrigues and Pinto 2022).

The process of extracting useful predictions from unorganized social media content proves to be a recurring challenge. The study by Wunderlich and Memmert (2022) demonstrated that Twitter information together with in-play events and betting odds failed to surpass pre-game odds which resulted in no additional predictive value. The model developed by Beal et al. (2020) faced a limitation because it depended on a single media source which restricted its ability to work with diverse data sets. The research indicates that models struggle with robustness because they use limited and prejudiced datasets while hybrid approaches that combine multiple data sources still tend to perform worse than bookmaker predictions (Baboota and Kaur 2019; Elmiligi and Saad 2022).

However, there are several great potential in hybrid models. This way, hybrid models, which are built on a number of machine learning techniques, including classification and clustering algorithms, produce better prediction accuracies than conventional methods. For instance, it has been shown that hybrid methods outperform sole classification algorithms in that they can take advantage of the strengths of various approaches (Kinalioğlu and Kuş 2023). There is one more significant benefit that comes with using hybrid models: The ability to include new data sources such as social media sentiment and fan opinions in the predictive models. Integrating this extra information expands the data set to include less quantitative, but still important, information about the emotional context of the events, which may help to improve the predictive power (Kinalioğlu and Kuş 2023).

In addition, the hybrid models can help in the application of feature engineering techniques. These models receive input from past match statistics, player metrics, and team performance trends, which they use to enhance their prediction models and specialize them for particular contexts, such as a league or a tournament (Elmiligi and Saad 2022). The modular structure of hybrid models also enables researchers to develop new models that can be specific to a certain kind of match or environment and create new directions in the field of predictive analytics. For example, hybrid strategies have been used to analyze information from different leagues and therefore generate more accurate and comprehensive results when applying them in different settings (Rodrigues and Pinto 2022).

The hybrid models enable additional possibilities for application. Yeung, Bunker, and Fujii (2023) proved that using FIFA player ratings together with coaches' tactical choices produced better predictive results and provided useful information for team roster decisions, tactical strategy development and transfer market evaluation. The research of Razali et al. (2022) demonstrated that combining structured pi-ratings with deep learning methods resulted in better performance and RPS outcomes which indicates hybrid approaches can surpass conventional models. Zheng (2023) demonstrated that multimodal deep learning systems with stacking integration methods enable the combination of different features to improve both prediction accuracy and stability.

In conclusion, although the application of hybrid models is accompanied by such problems as increased computational complexity, scalability issues, and the nature of football matches, they also present a number of advantages. These include data source diversity, analytical methods, and context-specific predictions. These strengths make hybrid models a valuable instrument for enhancing the precision and coverage of football match outcome forecasts.

2.6 Summary

Chapter 2 offered a thorough discussion of the basic ideas and methods that are connected to this research. It began by outlining the importance of machine learning and sentiment analysis in predictive modeling of football match results. The discussion was then extended to hybrid models, which use both structured and unstructured data to increase the predictive accuracy. The key approaches, challenges, and evaluation metrics for hybrid models were then reviewed with their strengths and weaknesses highlighted. Furthermore, the chapter outlined the ability of hybrid models to diversify data sources and analytical methods to innovate within sports analytics. Furthermore, a gap in the literature was identified, particularly regarding the application of hybrid models that combine structured data with unstructured data, specifically sentiment analysis, in football match predictions over the past five years. To this end, a thorough search was made before the systematic literature review was conducted using PRISMA. This initial attempt was made to include papers from other fields that could provide useful knowledge and methods that could be applied to hybrid modelling of football outcome prediction. Therefore, the presented elements made it possible to proceed to the analysis of the problem and its research questions in a systematic way.

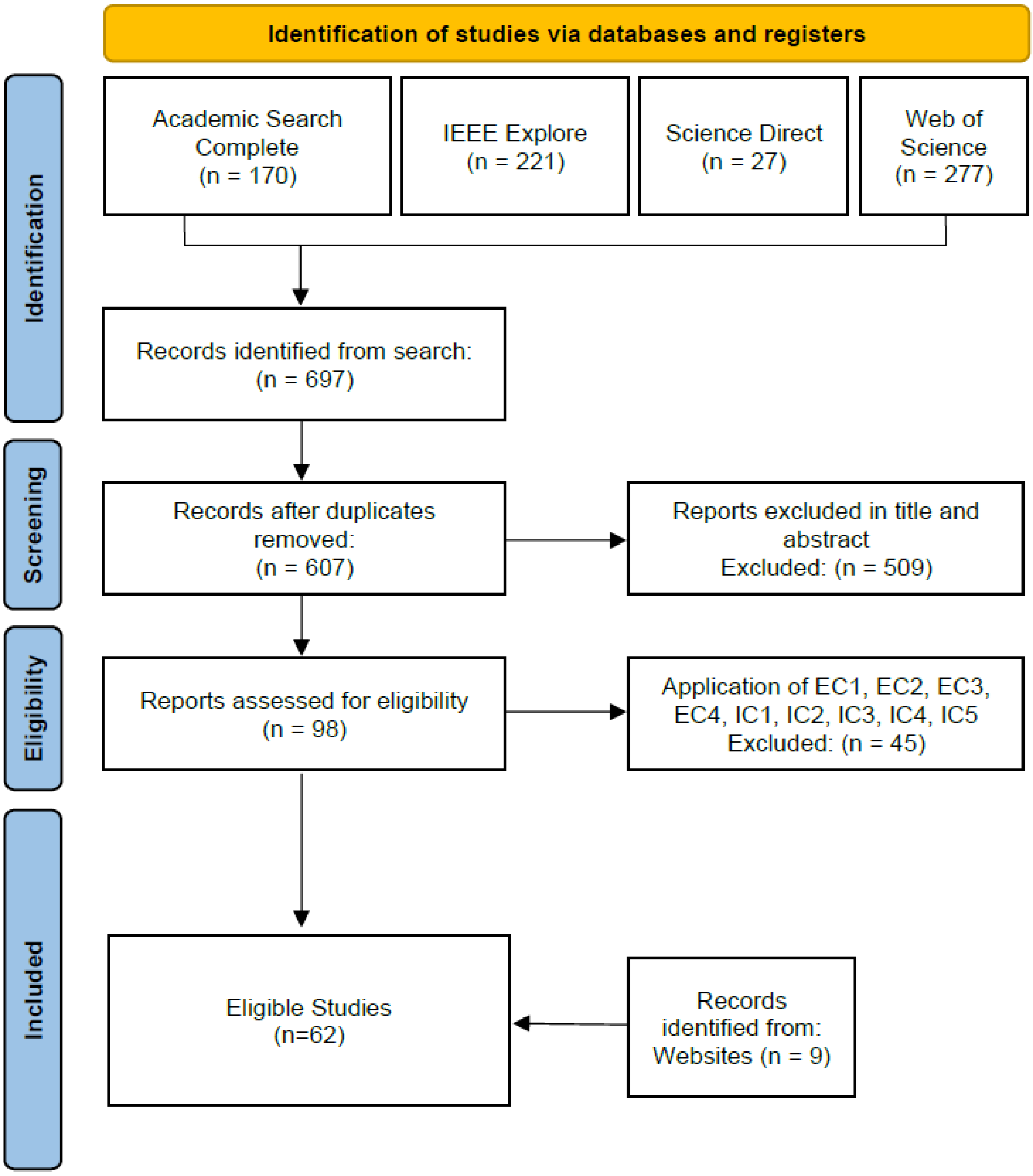


Figure 2.1: Prisma flow diagram.

Chapter 3

Methods and Materials

3.1 Introduction

Chapter 3 presents the methods and materials utilized in this study, outlining the datasets, tools, and techniques employed to develop and validate the predictive models. It highlights the preprocessing and integration of structured and unstructured data while addressing key data protection and privacy considerations. Additionally, the chapter provides an overview of the experimentation process, including an initial test with structured and sentiment-based data, laying the groundwork for more advanced hybrid modeling and comprehensive validation in subsequent phases.

3.2 Methods and Tools

This section presents the methodological choices and tools employed in the study. It details the extraction and preparation of data, exploratory analysis, feature engineering, the selection of models and algorithms, and the computational infrastructure used.

3.2.1 Data Extraction

Several works on football outcome prediction have relied on the freely available football-data.co.uk[1] dataset. However, upon closer inspection, this dataset contains mainly betting odds. The literature shows that betting odds data remains popular because it is easy to access but it introduces biases because bookmakers and bettors influence these odds more than actual match play. The presence of these biases in the data leads to inaccurate predictions and makes the results less reliable(Goto and Yamada 2023; Hegarty and Whelan 2025). Although this dataset is often used in the literature due to the difficulty of obtaining richer datasets and the need for rapid prototyping.

To mitigate this issue, in this study only the statistical match variables from football-data.co.uk were retained, and they were complemented with additional match statistics retrieved from API-Football[2] and Footystats[3]. Custom Python clients were developed to process the Comma-Separated Values (CSV) files and to consume the Application Programming Interface (API), with the resulting data consolidated into a PostgreSQL database. The final dataset covered 1900 EPL matches across five seasons (2020/2021 to 2024/2025).

[1] https://www.football-data.co.uk/englandm.php
[2] https://www.api-football.com
[3] https://footystats.org

The extraction pipeline operated through Python scripts (designed to be orchestrated as Directed Acyclic Graphs (DAGs) in Apache Airflow) to achieve both reproducibility and scalability. The quality control system operated to preserve match and source consistency through verification of results and correction of detected discrepancies.

To capture supporter opinion, tweets were collected from the X platform (formerly Twitter) using a custom Python script and the official (API, combined with X advanced search queries. The team-specific search queries for Premier League teams included keywords, hashtags, mentions and official team accounts to find relevant match-related tweets. The analysis considered only public tweets, applying several filters to ensure data quality. These included removing retweets and duplicates, restricting content to English, excluding private or unrelated messages, and retaining only tweets directly related to the matches. In addition, to reduce noise from pre-match hype or immediate reactions, only tweets posted within the 48 hours preceding each fixture were considered, explicitly excluding those published on the match day itself (Wunderlich and Memmert 2022). The research spanned the same five-year time frame from 2020/2021 through 2024/2025.

The metadata stored included the tweet_id and text, as well as public engagement indicators (retweet_count, reply_count, bookmark_count, impression_count, and like_count). To safeguard privacy, no usernames or geolocation data were retained.

3.2.2 Data Preparation

The structured data required multiple preprocessing operations to achieve complete data quality and modeling readiness. The dataset underwent an initial assessment to detect duplicate entries, missing information and data type inconsistencies. The removal of duplicate rows followed by necessary data type conversions for numerical features ensured all sources had compatible data types.

Systematic methods were applied to address missing data points. Features with more than 20% missing data were excluded, as such variables could introduce noise and bias into the results. Median imputation was applied to the home_offsides and away_offsides variables, since this method produces stable estimates that are less sensitive to extreme values. Zero imputation was applied to performance indicators, including goalkeeper saves, yellow/red cards by time intervals, and corner statistics, as a zero value indicates the absence of the recorded event.

The final inspection revealed that all missing data points received proper treatment. The datasets were stored in a structured format following cleaning and imputation procedures to maintain data consistency between matches and seasons.

The unstructured data consisting of tweets required multiple stages of cleaning and preprocessing to achieve text consistency and extract relevant features. The database retrieved raw tweets which then entered a deterministic processing system. The text processing system performed HyperText Markup Language (HTML) unescaping, Unicode normalization, Uniform Resource Locator (URL) placeholder replacement, punctuation standardization and whitespace compression to create a neutral text format which maintained team mentions and hashtags.

To capture stylistic and contextual signals, several features were derived from the raw text, such as counts of exclamation, question marks, ratio of uppercase tokens, word and character counts, as well as the presence of elongated words, emojis, and hashtags. A predefined lexicon enabled the calculation of emoji sentiment scores which added affective data to the analysis.

The system produced standardized text information about URLs, mentions, user interaction metrics and writing style indicators. The processed outputs were consolidated into tabular form, where each record linked the cleaned tweet text to its associated features and metadata, making the dataset suitable for downstream sentiment analysis and predictive modelling.

3.2.3 Data Exploratory Analysis

An exploratory analysis was conducted to examine how the data elements were distributed across the structured dataset and how they related to each other, while also identifying potential duplicate information. The analysis began with correlation studies among numerical variables, including both match statistics like goals, shots, possession and derived outcome indicators. The analysis of Full-Time Result (FTR) as a target variable required converting it into numerical values (H = 2, D = 1, A = 0) to perform correlation tests with independent features.

The analysis revealed strong correlations between different features which showed that certain statistical indicators duplicated each other. To mitigate multicollinearity, redundant variables with correlation coefficients exceeding a predefined threshold were flagged for removal. The analysis reduced the number of features in the dataset by keeping only the most important variables.

In addition to the correlation analysis, seasonal patterns were examined by calculating aggregate statistics per season, including goals, Expected Goals (xG), shots, fouls, yellow and red cards, and passing accuracy. The analysis of Premier League match patterns between 2020/2021 and 2024/2025 seasons revealed important trends through these statistical measurements.

The exploratory analysis created a solid foundation for feature engineering and model development by removing redundant data and establishing domain-specific descriptive statistics.

3.2.4 Feature Engineering

The dataset was enhanced through feature engineering to add contextual and derived variables which would improve predictive accuracy. The dataset received multiple transformation operations. First, categorical interactions were created to capture fixture-specific contexts, which included a matchup identifier for each pair of teams, a team pair feature independent of home–away order, and combinations linking the home team with the referee. Temporal indicators were then added by extracting the month of the match, the day of the week, and flags identifying games played in December, during winter or summer, or midweek. Additional flags were used to mark the start and end of a season by analysing the match order of each team, and kick-off times were grouped into morning, afternoon, and evening buckets. The models received comparative data through relative performance features were also engineered by computing ratios and differences between home and away team statistics, such as goals, shots, and cards, allowing the models to learn from comparative rather than absolute measures.

Composite indices were developed to evaluate both the strategic and behavioral aspects of team play. The offensive pressure index combined three team statistics which included shots on target, attacks, dangerous attacks while also evaluating home and away pressure levels through a pressure ratio. The physical aggressiveness measures derived from fouls and disciplinary records (yellow and red cards) and offensive efficiency ratios calculated the finishing accuracy through the comparison of shots on target to total shots. The fast start indicator used the first ten minutes of the match to combine goals with corners and cards as indicators of early match intensity.

Finally, half-time indicators to measure how goals, cards and corners distribute between the first and second halves of matches. The engineered features transformed the basic statistical data into a more detailed representation that included situational elements, temporal aspects, relational data to provide predictive models with enhanced match dynamic understanding.

3.2.5 Datasets

Following the stages of data preparation and feature engineering, several datasets were constructed to support the experimental design of this study. For the structured statistical component, different temporal partitioning strategies were applied in order to evaluate model performance under realistic forecasting conditions.

A series of rolling window datasets were generated with window sizes ranging from one to seven. In each case, the training set consisted of the first three seasons (2020/2021-2022/2023), the validation set corresponded to the fourth season (2023/2024), and the final test set was the 2024/2025 season. This approach allowed the assessment of predictive stability across different temporal contexts.

In parallel, an expanding window approach was implemented to progressively enlarge the training set over time. Four datasets were produced, the first trained exclusively on season 2020/2021 and validated/tested on 2021/2022, the second trained on the first two seasons with validation/testing on 2022/2023, the third trained on the first three seasons with validation/testing on 2023/2024, and finally, the fourth trained on the first four seasons with the last season (2024/2025) reserved for validation and testing. This design ensured that models were always evaluated on unseen future data, while capturing the cumulative learning effect of historical information.

Additionally, a time-series split strategy was applied, where seasons 2020/2021 to 2023/2024 were used jointly for training and validation via TimeSeriesSplit cross-validation, while the final season 2024/2025 was held out as a completely unseen test set. This method provided a rigorous framework for hyperparameter tuning while preserving temporal causality.

Alongside structured statistics, multiple tweet-derived datasets were also generated, reflecting the parallel line of investigation into social media sentiment and textual signals. These included tabular datasets constructed from sentiment analysis features extracted using Natural Language Toolkit (NLTK) Valence Aware Dictionary for sEntiment Reasoning (VADER) and CardiffNLP's Twitter-roBERTa-base model, capturing polarity and intensity of supporter opinions. A separate dataset was prepared for training models with Term Frequency-Inverse Document Frequency (TF-IDF) representations of concatenated tweets, while additional datasets were generated from embedding models (BERTweet, Sentence-BERT, and Masked and Permuted Pre-training for Language Understanding (MPNet)), where each fixture was represented by aggregated vector embeddings of tweets associated with both teams.

Together, these datasets form the foundation for the hybrid modelling approach explored in this study, enabling a systematic comparison between purely statistical models, text-based models, and combined approaches that integrate both structured and unstructured sources of information.

3.2.6 Models and Algorithms

The research employed multiple machine learning algorithms to achieve diverse learning approaches and create an effective comparison system. The structured statistical datasets received models from multiple supervised learning categories. The linear baseline model used Logistic Regression but k-Nearest Neighbors (kNN) provided a non-parametric approach. The Multi-Layer Perceptron (MLP) neural network model received evaluation while SVM with radial basis function kernels detected non-linear decision boundaries. The research analyzed four ensemble-based methods which included Natural Gradient Boosting (NGBoost), XGBoost, Light Gradient-Boosting Machine (LightGBM) and Categorical Boosting (CatBoost) because these gradient boosting frameworks are known for their effectiveness in tabular prediction tasks.

The tabular sentiment datasets obtained from VADER and CardiffNLP outputs received analysis through both Logistic Regression and XGBoost models. Logistic Regression was used as a strong linear baseline, whereas XGBoost was selected for its ability to model complex non-linear relationships and feature interactions in structured data.

The primary algorithm used for TF-IDF representations of concatenated tweets was Logistic Regression. The selection of this algorithm stemmed from its established success in processing high-dimensional sparse feature spaces which is a standard method for text classification tasks.

The baseline classifier used for embedding-based representations (BERTweet, Sentence-BERT and MPNet) was Logistic Regression because of its stable performance and accurate probability generation. The evaluation of dense semantic vector spaces included Linear SVM testing alongside the most successful embedding models.

The study investigated hybrid approach potential through ensemble methods which included soft voting, hard voting and stacking to combine predictions from separate models trained on structured match statistics and tweet-derived datasets. This provided a framework to assess whether combining heterogeneous information sources could enhance predictive accuracy beyond that of individual models.

3.2.7 Tools and Computational Infrastructure

All experiments were conducted in Python 3.11, taking advantage of its rich ecosystem for data science and machine learning. The research used Pandas and NumPy for data management and scikit-learn for traditional machine learning while XGBoost, LightGBM, CatBoost and NGBoost handled gradient boosting tasks, while NLTK (VADER) and Hugging Face Transformers performed sentiment analysis and text embedding operations. Matplotlib and Seaborn provided support for visualization work and exploratory data analysis. The PostgreSQL relational database managed data storage and integration between statistical data and tweet-derived information through its structured database system.

The experimental environment consisted of a laptop running Ubuntu 24.04, with an Intel(R) Core(TM) i7-10750H Central Processing Unit (CPU) @ 2.60GHz processor, 32 GB of RAM and an NVIDIA RTX 2070 Mobile Graphics Processing Unit (GPU) for transformer model training and inference operations. The virtualenv environment management system maintained dependency stability while ensuring experiment reproducibility through its dependency control features.

3.3 Experimentation and Validation

The experimentation stage served to create reliable reference points and confirm the research design before moving on to more sophisticated models. Logistic Regression was selected as the baseline model for structured statistical data due to its clear interpretation and proven success in comparable classification problems. The initial experiments were conducted exclusively on the simple rolling window datasets, where the training set consisted of the first three seasons (2020/2021-2022/2023), the validation set corresponded to the fourth season (2023/2024), and the final season (2024/2025) was reserved as the test set. This setup ensured that the evaluation respected the natural chronological order of the data while providing a robust benchmark for subsequent model development.

Complementary experiments were also conducted by increasing the training data size through seasonal additions while using the following season for validation and testing purposes. TimeSeriesSplit served as a cross-validation method for both rolling window and expanding window datasets to provide detailed evaluation while maintaining the natural sequence of time-dependent data. The final testing of all setups used the 2024/2025 season as a dedicated holdout set to obtain performance results on future data that had not been seen before.

For the tweet-derived tabular datasets, Logistic Regression was again used as the baseline, through two stages using VADER sentiment features followed by CardiffNLP Twitter-roBERTa-base model results. XGBoost was applied to the CardiffNLP dataset to detect complex relationships between sentiment features. Given the high number of variables in these tabular datasets, feature selection methods were employed to retain only the most informative features prior to training.

The evaluation process employed multiple assessment metrics which worked together as a single evaluation system. Log Loss was defined as the primary metric, as it directly assesses predicted probability quality and is sensitive to poor calibration. Additional metrics included Brier Score, Balanced Accuracy, RPS, Macro F1, providing complementary views of probability calibration, class balance, and overall performance across the three outcome categories. To ensure reproducibility and comparability of results, fixed random seeds were maintained throughout all experiments.

3.4 Data Protection and Privacy

All data used in this study were collected through procedures that complied with data protection standards. The X API, together with advanced search queries, was used to retrieve publicly available tweets for the social media analysis. No personal details such as usernames, user IDs, locations, or device information were stored. Instead, tweet IDs, textual content, and engagement metrics were retained, with the analysis focused on aggregated insights to minimise data collection and prevent individual profiling. The processing of publicly available social media data was carried out solely for academic research purposes, in line with the principle of legitimate interest under data protection regulations. No sensitive personal data, such as political or religious opinions, were collected or analysed.

The match statistics data originated from public and licensed providers that contained no information that could be classified as personal or sensitive. All datasets were stored within a PostgreSQL environment running on a protected laptop, following secure practices that included restricted access, regular dependency updates, and Hypertext Transfer Protocol Secure HTTPS-based API connections.

3.5 Summary

This chapter detailed the methods and materials employed in this study, beginning with a comprehensive discussion of the datasets, including structured match statistics from football-data.co.uk, API-Football, and Footystats, as well as sentiment features derived from tweets collected via the X API. The tools and techniques for preprocessing, integration, and analysis were outlined, while placing strong emphasis on data protection and privacy measures to ensure full compliance with standards such as the General Data Protection Regulation (GDPR). The experimentation phase introduced preliminary baseline tests using Logistic Regression on structured data with rolling windows, establishing a reference framework for evaluation. This foundation sets the stage for subsequent experimentation with unstructured text data and hybrid models, aimed at enhancing predictive accuracy and generating deeper insights into football match outcomes.

Chapter 4

Implementation, Analysis and Results Discussion

4.1 Introduction

The fourth chapter presents the implementation of the predictive models together with their evaluation outcomes based on the research design established in Chapter 3. In line with the DSR methodology, this stage reflects the construction and assessment of the artefacts developed to address the research problem. The chapter details the implementation of the models, the datasets on which they were trained, and the validation procedures employed. The results are organised according to the type of data used across structured match statistics, tweet-derived datasets, and hybrid approaches, progressing from baseline models to more advanced techniques. Each section presents the outcomes of the experiments, followed by an analysis and discussion of the findings, highlighting both the predictive performance achieved and the broader insights gained.

4.2 Implementation

The baseline experiments focused on models trained exclusively with structured match statistics, provided a reference framework for subsequent comparisons with text-based and hybrid approaches. All baseline models were implemented using logistic regression within a consistent pipeline that applied standard preprocessing procedures, including feature scaling, categorical encoding, and imputation where required. The input datasets were loaded from the structured files prepared in the data processing stage, ensuring that identical features were used across all temporal validation strategies.

To evaluate the models, three different temporal partitioning strategies were tested. The first was the rolling window, where the model was trained on three consecutive seasons and validated on the following one, with the 2024/2025 season kept exclusively for final out-of-sample testing. The second was the expanding window, which started training on the earliest season and gradually incorporated additional seasons, while validation and testing were always performed on the immediately subsequent season. Finally, a TimeSeriesSplit approach was used, where the first four seasons were internally divided into multiple train-validation folds, and the last season was consistently held out for testing.

By directly contrasting these configurations across identical datasets and preprocessing pipelines, it was possible to assess their suitability for football outcome prediction. Based on these experiments, the TimeSeriesSplit with rolling windows was selected as the standard evaluation method. This choice balanced the need for realistic temporal forecasting with efficient use of historical data. The resulting configuration was then consistently adopted in all subsequent experiments, ensuring methodological comparability between models trained on different sources of information.

Following this decision, a broad set of machine learning algorithms was trained using structured statistical features. The models covered a wide spectrum of approaches, ranging from ensemble-based methods (Random Forest, XGBoost, LightGBM, CatBoost, NGBoost) to distance and kernel-based classifiers (kNN, SVM with RBF kernel), as well as neural architectures (MLP). Regardless of the algorithm, each model was integrated into a uniform pipeline that standardised the experimental process. This pipeline consistently applied preprocessing steps such as scaling of numerical attributes, one-hot encoding of categorical features, and imputation of missing values where necessary, ensuring that any performance differences could be attributed to the learning algorithm itself rather than inconsistencies in data preparation.

The model training relied on careful hyperparameter selection, carried out either through grid search or via library-specific optimisation routines, depending on the algorithm. To prevent overfitting and reduce computational overhead, early stopping was enabled for all gradient-boosting models (XGBoost, LightGBM, CatBoost, NGBoost). For algorithms particularly sensitive to feature scaling, such as SVM and kNN, systematic standardisation was applied within the pipeline. Finally, all experiments were executed with fixed random seeds, ensuring that results remained fully reproducible.

The training and evaluation procedure followed the same sequence across models, beginning with the partitioning of datasets using the selected rolling-window TimeSeriesSplit configuration. The models were then trained on the training folds with the chosen preprocessing, after which validation on the held-out folds was used to tune hyperparameters and assess generalisation. Finally, the models were retrained on the entire training period and evaluated on the 2024/2025 holdout season. By adhering to this unified setup, it was possible to build a coherent benchmark of statistical models and ensure direct comparability with subsequent experiments that integrated text-derived features.

After establishing the benchmarks with statistical features, attention turned to models built on tweet-derived information. In this stage of the implementation, the baseline experiments with tweet-derived features were carried out. Two sentiment-based datasets were considered, one constructed from NLTK VADER outputs and the other from CardiffNLP's Twitter-RoBERTa model. Both datasets contained fixture-level aggregated features together with the full-time result (FTR) as the prediction target.

To ensure comparability across approaches, the same preprocessing pipeline was applied in both cases. A whitelist of variables was used to avoid information leakage and redundancy, categorical variables were one-hot encoded, and numerical variables were standardised. The order of the prediction classes was fixed as [A, D, H] to guarantee consistent evaluation across experiments.

During the baseline phase, all temporal partitioning strategies described in Chapter 3 were tested. Based on these comparisons, TimeSeriesSplit with rolling windows was selected as the standard configuration, and this choice was applied to all subsequent experiments. In every case, the 2024/2025 season was kept as a holdout set, ensuring that the final evaluation was always performed on unseen data.

The initial baseline models were Logistic Regression classifiers trained separately on the VADER and CardiffNLP datasets, following a unified pipeline with fixed random seeds to guarantee reproducibility.

In parallel, the same datasets were also used to train XGBoost classifiers, where early stopping was applied with a validation split and a focused hyperparameter search was carried out with Log Loss as the optimisation target. After selecting the best configuration, the model was retrained using both the training and validation sets before being evaluated on the holdout season. As with Logistic Regression, the number and type of features generated after one-hot encoding were carefully tracked to ensure transparency and reproducibility.

Finally, in order to maintain a fair comparison across models, all experiments shared the same evaluation routine and metrics. The standardised pipeline was applied to both the VADER and CardiffNLP datasets, as well as to Logistic Regression and XGBoost, ensuring that performance differences could be attributed to the models and data representations themselves rather than to implementation inconsistencies.

After completing the sentiment-based experiments, the next step was to explore TF-IDF representations of tweets. To do so, a dedicated dataset was created in which all tweets related to a fixture were concatenated by team and transformed into sparse vector representations. To better understand the impact of text preprocessing, four variants of this dataset were prepared: a raw version, one with stop word removal, one with stemming, and another with lemmatisation.

Each version was then split according to the temporal scheme established earlier, with the first four seasons used for training and validation and the 2024/2025 season kept as a strict holdout for final testing. This ensured that the evaluation consistently respected chronological order. The implementation relied on scikit-learn's TFIDFVectorizer, applied separately to the tweets associated with home and away teams. While configuring the vectoriser, parameters such as minimum and maximum document frequency, n-gram range, and maximum feature size were tuned to balance representation richness against sparsity. After the vectorisation step, the resulting feature matrices were integrated with team identifiers through a preprocessing pipeline.

To address the classification task, the experiments relied on Logistic Regression, a model well established for its effectiveness in high-dimensional sparse spaces. The saga solver was employed to cope with the large number of features, while the regularisation strength was tuned through hyperparameter search. All training pipelines combined preprocessing and classification into a single reproducible workflow, with fixed random seeds applied throughout to ensure consistency. In each experimental run, the model was first trained on the training folds of the TimeSeriesSplit, then validated on sequential folds for hyperparameter refinement, and finally retrained on the combined training and validation data before being evaluated on the 2024/2025 holdout season.

This design provided a consistent framework in which the impact of different preprocessing strategies (raw text, stop word removal, stemming, and lemmatisation) could be directly compared. By maintaining the same pipeline structure and evaluation metrics, it ensured that any observed differences could be attributed to the preprocessing approach rather than inconsistencies in implementation.

The final stage of the tweet-based approach focused on models trained with dense semantic embeddings. Each fixture was represented by aggregated embeddings of all tweets associated with both teams, producing compact numerical vectors that capture contextual meaning beyond simple word frequency.

The first set of experiments relied on Sentence-Bidirectional Encoder Representations from Transformers (BERT) with the MPNet architecture, which is widely recognised for its strong performance in semantic similarity tasks, and on top of these embeddings two classifiers were tested. Logistic Regression served as the baseline because of its reliable probability estimates, while a Linear Support Vector Machine was included to assess whether a margin-based classifier could extract additional performance from the dense feature space. To ensure comparability, both models were trained with the same temporal splits as before, maintaining consistency with earlier experiments.

Another variation of Sentence-BERT (SBERT) was evaluated using the Minimal Language Model (MiniLM) architecture, which produces more lightweight embeddings while still preserving semantic richness. The Logistic Regression classifier was again employed, ensuring a consistent point of comparison across embedding families.

In addition, the study explored BERTweet, a model pre-trained specifically on Twitter data. Its embeddings are tailored to capture linguistic nuances such as hashtags, mentions, and informal expressions that frequently appear in football-related tweets. The Logistic Regression classifier was chosen here as well, keeping the evaluation framework aligned with the other embedding-based experiments.

Finally, FastText with 300-dimensional embeddings was used to generate document-level representations by averaging word vectors. This approach offered a simpler yet competitive baseline for comparison with transformer-based embeddings. The Logistic Regression classifier was selected, reflecting its stable behaviour in high-dimensional but relatively low-noise embedding spaces.

All embedding-based models followed the pipeline structure and validation routine established earlier. The embedding-based models were trained on sequential folds using TimeSeriesSplit, with hyperparameters adjusted within a narrow range, and the final evaluation consistently carried out on the unseen 2024/2025 season. By applying this consistent framework, the experiments enabled a fair comparison of different embedding strategies and classifiers, highlighting the contribution of semantic representations to football outcome prediction.

After completing the standalone experiments with statistical and tweet-based models, the focus shifted to hybrid approaches that integrated both sources of information. The aim was to investigate whether combining structured match statistics with sentiment-derived features could improve predictive performance compared to using either source in isolation.

Several ensemble methods were implemented to combine the strengths of individual classifiers. The first approach relied on soft voting, where probability outputs from independent models were aggregated into a joint prediction. Within this pipeline, both global weighting schemes and class-specific weighting were explored, allowing the relative contribution of each model to vary depending on the outcome category. In addition, a dynamic voting strategy was tested, in which the weights assigned to models were adapted according to their local performance in validation folds.

In addition to voting ensembles, a stacking strategy was also developed. Here, the probability outputs from the base models were fed into a Logistic Regression meta-learner trained to optimise the final prediction. Building on this idea, a further variation introduced a residual learning scheme, conceptually framed as a special case of stacking. In this configuration, the statistical model acted as the base learner, producing probability distributions over match outcomes, while the gap between these baseline probabilities and the observed outcomes was encoded as one-hot vectors and treated as the residual signal. A lightweight and regularised Logistic Regression was then adopted as the meta-model, trained on tweet-derived features to approximate these residuals. At inference time, the correction estimated by the residual model was added to the original statistical probabilities and renormalised to yield a valid probability distribution. In this way, the sentiment-based component did not operate as an independent predictor but rather as an adjustment mechanism that systematically corrected biases in the statistical baseline, making use of the complementary signal in tweets to refine statistical predictions.

Alongside the ensemble methods, an early fusion strategy was explored. In this approach, sentiment features derived from VADER analysis were directly concatenated with the structured statistical features to create an extended tabular dataset. This design enabled a Logistic Regression classifier to learn from both sources of information simultaneously, which is characteristic of early fusion as the integration happens at the feature level rather than only at the decision stage.

All hybrid approaches followed the same experimental procedure as the standalone models. For consistency, temporal splits were applied using the rolling-window TimeSeriesSplit configuration, and the 2024/2025 season was kept strictly as a holdout set.

Together, these implementations established a coherent and reproducible foundation on which the comparative analysis of results could be reliably conducted.

4.3 Evaluation Metrics

The evaluation of football outcome prediction models in this study relied on a combination of probabilistic and classification-based metrics. The goal was not only to identify the most likely result of a match, but also to provide well-calibrated probability distributions across the three possible outcomes: home win (H), draw (D), and away win (A). Accordingly, the chosen metrics capture both classification performance and probability calibration.

Log Loss (Cross-Entropy Loss)

Log Loss was chosen as the primary evaluation metric, as it penalises models that are confident but wrong, providing a direct assessment of probabilistic calibration.

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i^{(H)}\log p_i^{(H)} + y_i^{(D)}\log p_i^{(D)} + y_i^{(A)}\log p_i^{(A)}\right] \tag{4.1}$$

where $p_i^{(H)}, p_i^{(D)}, p_i^{(A)}$ are the predicted probabilities for home win, draw, and away win in match i, and $y_i^{(H)}, y_i^{(D)}, y_i^{(A)}$ are binary indicators equal to 1 if the outcome occurred, 0 otherwise.

Example: If the model assigns a probability of 90% to a home win but the away team actually wins, the penalty is very high, reflecting an overconfident error.

Brier Score

The Brier Score measures the squared error between predicted probabilities and the actual outcomes, offering a more interpretable view of probability quality.

$$BS = \frac{1}{N}\sum_{i=1}^{N}\left[(p_i^{(H)} - y_i^{(H)})^2 + (p_i^{(D)} - y_i^{(D)})^2 + (p_i^{(A)} - y_i^{(A)})^2\right] \tag{4.2}$$

where $p_i^{(H)}, p_i^{(D)}, p_i^{(A)}$ are the predicted probabilities and $y_i^{(H)}, y_i^{(D)}, y_i^{(A)}$ are the observed outcomes (1 if occurred, 0 otherwise).

Example: If the model predicts 60% home, 30% draw, 10% away, and the match ends in a draw, the Brier Score penalises the squared distance between forecasted probabilities and the actual outcome.

Expected Calibration Error (ECE)

ECE directly measures the alignment between predicted confidence and observed frequencies by grouping predictions into bins and comparing average confidence with empirical accuracy, which provides an estimate of model calibration.

$$ECE = \sum_{m=1}^{M}\frac{|B_m|}{N}\left|\mathrm{Acc}(B_m) - \mathrm{Conf}(B_m)\right| \tag{4.3}$$

where B_m is the set of predictions falling into bin m, $\mathrm{Acc}(B_m)$ is the empirical accuracy in that bin, and $\mathrm{Conf}(B_m)$ is the average predicted confidence.

Example: If across 100 games the model predicted home win with 75% confidence but home teams only won 60 of them, the model is miscalibrated in that region, and the ECE captures this gap.

Accuracy

Accuracy measures the share of matches in which the most probable predicted outcome corresponds to the actual result.

$$\mathrm{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\Big(\arg\max\{p_i^{(H)}, p_i^{(D)}, p_i^{(A)}\} = \mathrm{TrueOutcome}_i\Big) \tag{4.4}$$

where $\mathbf{1}(\cdot)$ is the indicator function, equal to 1 if the predicted class matches the true outcome and 0 otherwise.

Example:

- Match 1: Prediction = Home (0.55), True = Home ⇒ Correct (1).
- Match 2: Prediction = Home (0.60), True = Draw ⇒ Incorrect (0).
- Match 3: Prediction = Away (0.52), True = Away ⇒ Correct (1).

Thus, Accuracy $= (1 + 0 + 1)/3 = 66.7\%$.

Balanced Accuracy

Balanced Accuracy computes the average recall across the three outcome categories, so that less frequent outcomes, such as draws, receive adequate consideration.

$$\text{Balanced Accuracy} = \frac{1}{3}\Big(\text{Recall}_H + \text{Recall}_D + \text{Recall}_A\Big) \tag{4.5}$$

where Recall_H, Recall_D, and Recall_A are the recalls for home win, draw, and away win respectively.

Example: If a model never predicts draws, its overall Accuracy might still be high, but Balanced Accuracy will drop significantly because the recall for the draw class is zero.

Macro F1 Score

The Macro F1 Score is the unweighted average of class-wise F1 values, giving equal importance to all outcome categories.

$$\text{Macro F1} = \frac{1}{3}\Big(F1_H + F1_D + F1_A\Big) \tag{4.6}$$

where $F1_H$, $F1_D$, and $F1_A$ are the F1 scores computed separately for home win, draw, and away win.

Example: A model that predicts home and away wins well but consistently fails to detect draws will achieve a low Macro F1, despite potentially high Accuracy.

Ranked Probability Score (RPS)

The RPS evaluates how close the predicted cumulative distributions are to the true outcome, rewarding predictions that respect the ordinal structure of football results, where a home win is closer to a draw than to an away win.

$$\text{RPS} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\sum_{k=1}^{2}\left(\sum_{c=1}^{k}(p_i^{(c)} - y_i^{(c)})\right)^2 \tag{4.7}$$

where outcomes are ordered as Away ($c = 1$), Draw ($c = 2$), and Home ($c = 3$). The inner sum accumulates the difference between predicted and observed cumulative probabilities up to class k.

Example: If the model predicts 45% home, 40% draw, 15% away and the match ends in a draw, the penalty is small because probabilities were close to the true result. By contrast, predicting 90% away win would yield a much larger error.

4.4 Results

This section presents the experimental results obtained with models trained on statistical match features, tweet-derived features, and their hybrid combinations. The results are organised into four parts for clarity: baseline comparisons, models trained on match statistics, models trained on tweets, and hybrid models. In terms of evaluation, performance is reported using multiple evaluation metrics, with the 2024/2025 season serving as a temporal holdout test set.

4.4.1 Baseline Models

Table 4.1: Baseline models configurations and test Log Loss

Configuration	Test Log Loss
Naive – Uniform (1/3, 1/3, 1/3)	1.0986
Naive – Empirical class prior (train 1–4)	1.0797
Rolling Window (train=1–3, val=4, test=5, ws=3)	0.9704
Expanding Window (train 1–4, val/test=5)	1.0333
Rolling Window + TimeSeriesSplit (train/val=1–4, test=5, ws=3)	0.9505
Expanding Window + TimeSeriesSplit (train/val=1–4, test=5)	1.0521
Chosen baseline: Rolling Window + TimeSeriesSplit (ws=3)	**0.9505**

The baseline experiments provided an initial point of reference for evaluating subsequent models, as summarised in Table 4.1. To this end, two naive predictors were included. The first was a uniform model that assigned equal probability to each outcome, resulting in a Log Loss of 1.0986. The second was an empirical class prior that reflected the distribution observed in the training set, with a Log Loss of 1.0797. Together, these predictors served as lower-bound expectations, illustrating the level of performance achievable without making use of match-specific information.

Beyond these trivial baselines, Logistic Regression was trained under several temporal partitioning schemes. In one setup, a simple rolling window using three seasons for training and one for validation achieved a Log Loss of 0.9704 on the holdout season. When the procedure was combined with TimeSeriesSplit for internal validation, the rolling window configuration reached 0.9505, while the expanding window dropped to 1.0521.

By contrast, the rolling window with TimeSeriesSplit delivered stronger results for Accuracy, Balanced Accuracy, Macro-F1, and RPS, while still maintaining a competitive Log Loss. Based on these findings, the Rolling Window + TimeSeriesSplit (Window Size (WS)=3) configuration was chosen as the statistical baseline, offering a balanced benchmark (Log Loss = 0.9505) against which tweet-based and hybrid models were later evaluated.

4.4.2 Models Based on Match Statistics

The experiments with statistical features were designed to assess the predictive power of models trained solely on structured match data. For each fixture, teams were represented by tabular features drawn from their three most recent matches played under the same home or away condition. This rolling-window approach allowed the models to capture short-term form while preserving the chronological order of the competition. Finally, the 2024/2025 season was kept aside as the holdout set for final evaluation.

Table 4.2 and Figure 4.1 summarise the performance of the models across a range of evaluation metrics. To evaluate predictive performance, Log Loss was used as the primary criterion, while additional measures such as ECE, Brier Score, Balanced Accuracy, Macro-F1, overall Accuracy, and the RPS offered complementary perspectives on predictive quality and calibration.

Within the tested approaches, Support Vector Machines with an RBF kernel achieved the lowest Log Loss on the holdout set (0.9066), closely followed by Logistic Regression with feature selection (0.9111). The gradient boosting ensembles such as XGBoost (0.9232), CatBoost (0.9266), and NGBoost (0.9362) also ranked among the strongest performers, delivering stable results across both discrimination and calibration metrics. In addition, Random Forests and LightGBM performed competitively, with Log Loss values of 0.9298 and 0.9328, respectively.

The additional analysis of the SVM (Radial Basis Function (RBF)) offered further insights into its behaviour across validation folds, out-of-fold predictions, and the holdout set. As illustrated in Figures 4.2, 4.3, 4.4, 4.5 and Table 4.3, the differences in Log Loss between validation, Out-Of-Fold (OOF), and test remained small. Moreover, the calibration plots aligned closely with the diagonal, indicating that the predicted probabilities matched well with the observed frequencies. Taken together, these results suggest that the model generalised consistently across evaluation splits.

By contrast, distance-based models such as kNN (0.9421) and neural models such as MLP (0.9413) achieved weaker results, showing higher Log Loss and lower Macro-F1 scores. Overall, the experiments indicated that tree-based ensembles and kernel-based classifiers were the most effective at extracting predictive signals from structured match statistics.

When comparing families of models, kernel-based methods and ensemble approaches consistently outperformed distance-based and neural models, while linear methods such as Logistic Regression remained competitive. The validation folds showed only moderate dispersion, with most Log Loss values falling between 0.88 and 0.96, as illustrated by the boxplot. Moreover, models that reached lower Log Loss also tended to achieve favourable Brier Score and RPS values, whereas overall Accuracy stayed within a narrow range of about 57% to 59% across most approaches.

Table 4.2: Performance of models trained on match statistics (WS=3) on the holdout 2024/25 season.

Model	Log Loss	ECE	Brier	Bal. Acc.	Macro-F1	Acc.	RPS
CatBoost	0.9266	0.5711	0.1819	52.28 %	0.4794	58.16 %	0.1209
kNN	0.9421	0.5359	0.1848	51.15 %	0.4758	57.11 %	0.1248
LightGBM	0.9328	0.5801	0.1822	50.84 %	0.4533	57.11 %	0.1202
Logistic Reg.	0.9111	0.5967	0.1791	50.55 %	0.4440	57.11 %	0.1184
MLP	0.9413	0.6082	0.1857	49.58 %	0.4266	56.32 %	0.1237
NGBoost	0.9362	0.5956	0.1828	52.53 %	0.4688	59.21 %	0.1202
Random Forest	0.9298	0.5551	0.1829	52.70 %	0.4886	58.68 %	0.1227
SVM (RBF)	0.9066	0.5804	0.1784	51.53 %	0.4471	58.16 %	0.1209
XGBoost	0.9232	0.5721	0.1806	53.14 %	0.4877	59.21 %	0.1201

4.4.3 Models Based on Match Tweets

When analysing the models trained on tweets, the goal was to assess how different textual representations could capture signals relevant to predicting match outcomes. These models

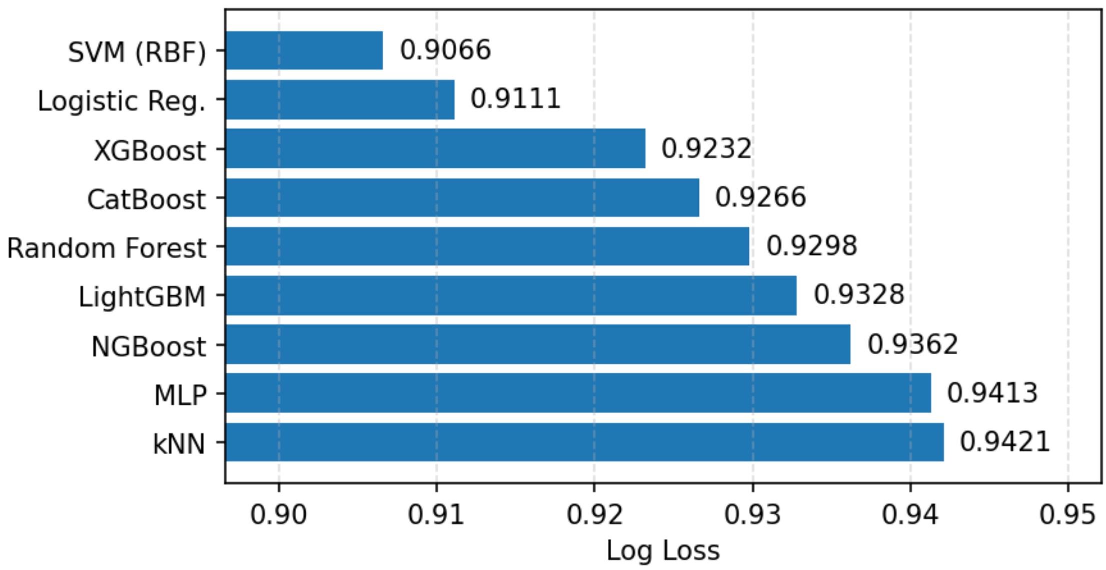


Figure 4.1: Test Log Loss for models trained on match statistics (WS=3).

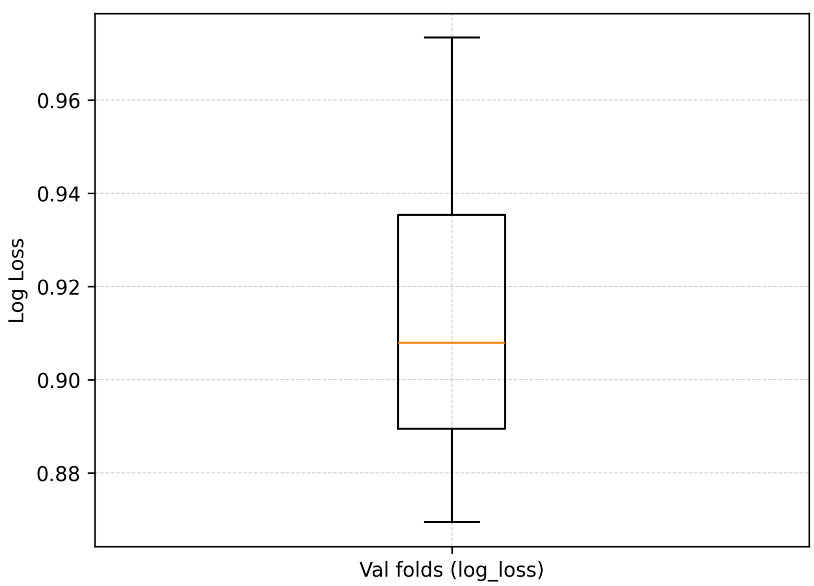


Figure 4.2: Validation fold dispersion for the SVM (RBF) model (Log Loss).

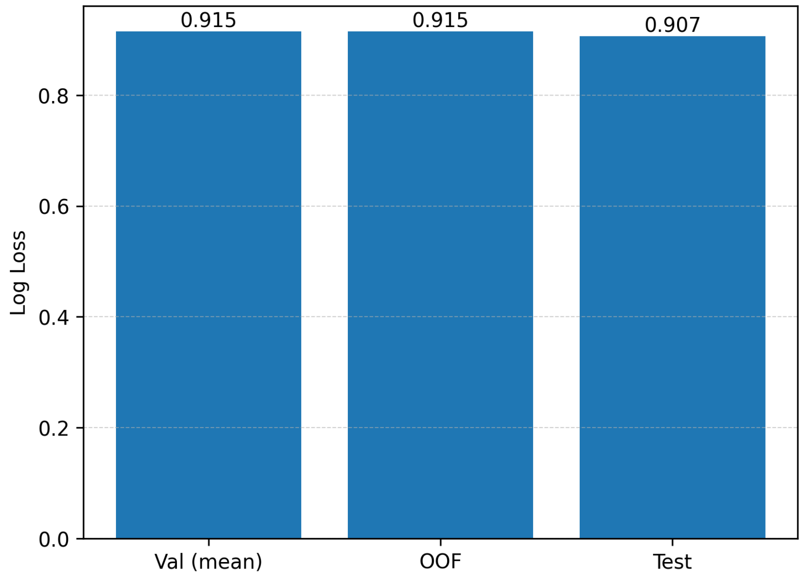


Figure 4.3: Comparison of Validation, OOF, and Test Log Loss for the SVM (RBF) model.

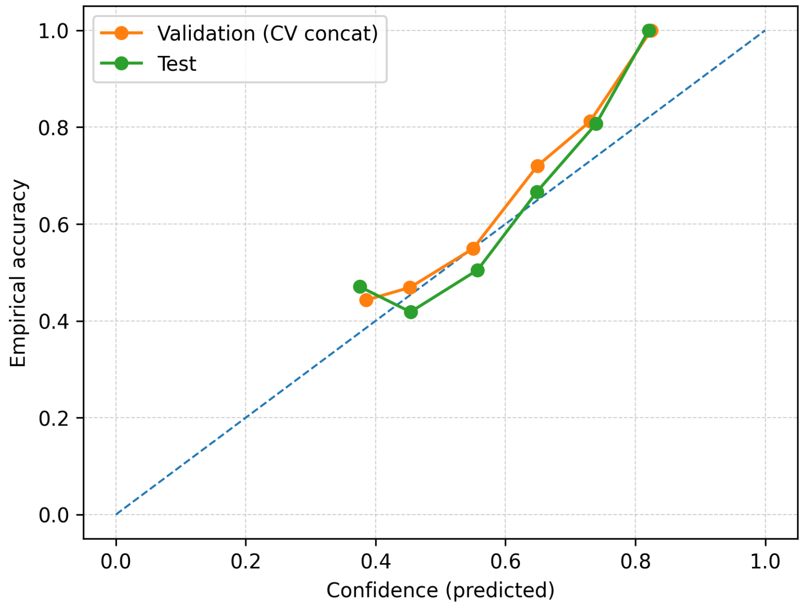


Figure 4.4: Reliability diagram comparing Validation (Cross-validation (CV) concat) and Test calibration for the SVM (RBF) model.

Table 4.3: Validation, OOF and Test metrics with corresponding deltas for the SVM (RBF) model.

Metric	Val (mean ± sd)	OOF	Test	Δ(Test–Val)	Δ(Test–OOF)
Log Loss	0.9151 ± 0.0406	0.9152	0.9066	-0.0085	-0.0086
RPS	0.1219 ± 0.0071	0.1219	0.1179	-0.0040	-0.0040
ECE	0.5612 ± 0.0227	0.5628	0.5804	+0.0192	+0.0176

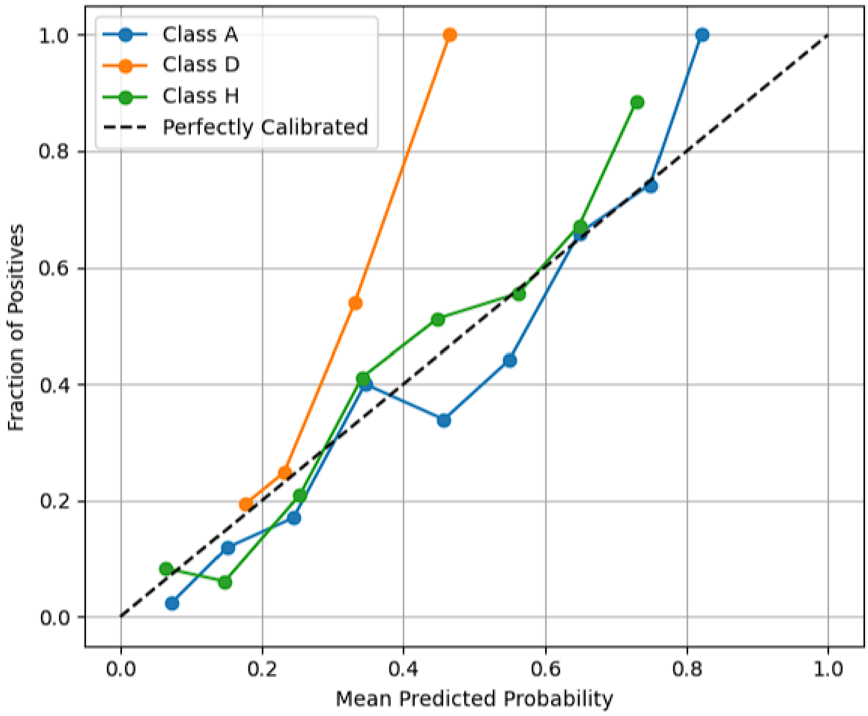


Figure 4.5: Calibration curves of predicted probabilities for the SVM (RBF) model.

were evaluated under the same temporal scheme used for the statistical approaches, with the 2024/2025 season kept aside as the holdout set for final testing.

As shown in Table 4.4, sentiment-based features provided an initial point of reference. Using these features, Logistic Regression with VADER reached a Log Loss of 1.0592, performing slightly better than CardiffNLP at 1.0618, and also achieved higher Accuracy and Macro-F1 values. When trained on the same features, XGBoost delivered results in a similar range, with both sentiment sources performing above the Logistic Regression baseline.

When shifting to TF-IDF representations, four preprocessing strategies were tested: raw text, stopword removal, stemming, and lemmatisation. The results were largely indistinguishable, with Log Loss values close to 1.05 and Accuracy hovering around 46 percent. However, none of the variants offered a clear improvement, and the Macro-F1 scores remained similarly aligned.

A broader range of outcomes was observed when embeddings were introduced. Among the

Logistic Regression baselines, SBERT-MPNet achieved the lowest Log Loss (1.0424), outperforming SBERT-MiniLM (1.0529), BERTweet (1.0463), and FastText (1.0637). Since SBERT-MPNet combined with Logistic Regression stood out within this group, it was further extended to a Linear SVM classifier. This additional experiment produced the best results across all tweet-based models, reaching a Log Loss of 1.0313 and an Accuracy of 47.89 percent, while Macro-F1 remained stable at 0.3558.

A detailed evaluation of the Linear SVM SBERT-MPNet model is provided in Table 4.6. In addition to its superior Log Loss and overall Accuracy (47.9 percent), the model achieved a Balanced Accuracy of 41.5 percent, a Brier Score of 0.2059, and an ECE of 0.5050. The RPS was 0.1444, while Macro-F1 remained consistent at 0.3558. As illustrated in Figure 4.6, the validation folds showed a narrow dispersion, with most Log Loss values falling between 0.97 and 1.04. The comparison between validation, OOF, and test performance in Figure 4.7 further indicates that the differences were small, with Log Loss varying by less than 0.03 across splits. Finally, the reliability diagram in Figure 4.8 demonstrate that validation and test calibration curves followed a similar trajectory, although deviations remained visible across outcome classes. The calibration curves in Figure 4.9 indicate a generally consistent alignment with the diagonal, while visible deviations persist across classes, most notably for class D.

Overall, the tweet-based models achieved Log Loss values between 1.0313 and 1.0755, with Accuracy ranging from 45.79 to 47.89 percent. Within this relatively narrow performance range, the Linear SVM with SBERT-MPNet consistently ranked highest and was therefore selected as the reference tweet-based model for subsequent comparisons, where the corresponding metrics and calibration details are reported in Table 4.6.

Table 4.4: Performance of tweet-based models (holdout 2024/25). Reported metrics: Log Loss, Accuracy, and Macro-F1.

Model	Log Loss	Accuracy	Macro-F1
LogReg Sentiment (Cardiff)	1.0618	46.32 %	0.3654
LogReg Sentiment (VADER)	1.0592	47.11 %	0.3878
XGBoost Sentiment (Cardiff)	1.0755	46.58 %	0.3650
XGBoost Sentiment (VADER)	1.0668	46.32 %	0.3839
LogReg TF-IDF (raw)	1.0502	46.32 %	0.3449
LogReg TF-IDF (stopwords)	1.0494	46.05 %	0.3418
LogReg TF-IDF (stemming)	1.0502	46.32 %	0.3449
LogReg TF-IDF (lemmatised)	1.0491	46.05 %	0.3419
LogReg SBERT (MiniLM-L6)	1.0529	45.79 %	0.3530
LogReg SBERT (MPNet)	1.0424	46.58 %	0.3531
LogReg BERTweet (base)	1.0463	45.79 %	0.3615
LogReg FastText (cc.en.300)	1.0637	47.37 %	0.3758
Linear SVM SBERT (MPNet)	**1.0313**	**47.89 %**	**0.3558**

4.4.4 Hybrid Models

The hybrid models combined statistical features with tweet-derived information to evaluate whether hybrid approaches could enhance predictive performance. For compatibility across sources, the statistical model based on an SVM with RBF kernel was adapted so that its

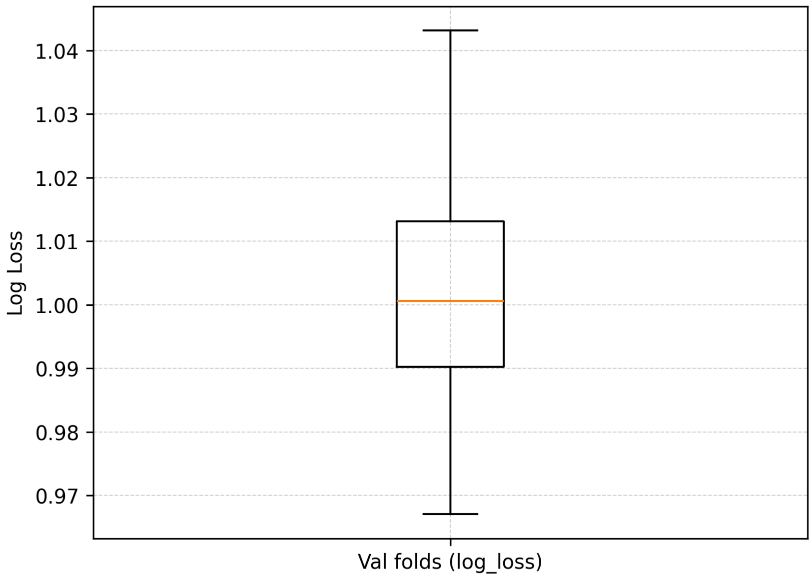


Figure 4.6: Validation fold dispersion for the Linear SVM SBERT (MPNet) model (Log Loss).

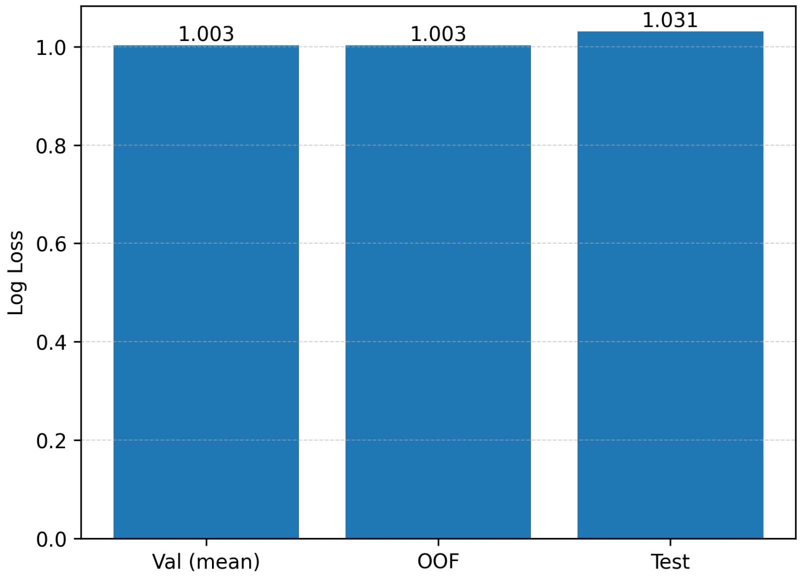


Figure 4.7: Comparison of Validation, OOF, and Test Log Loss for the Linear SVM SBERT (MPNet) model.

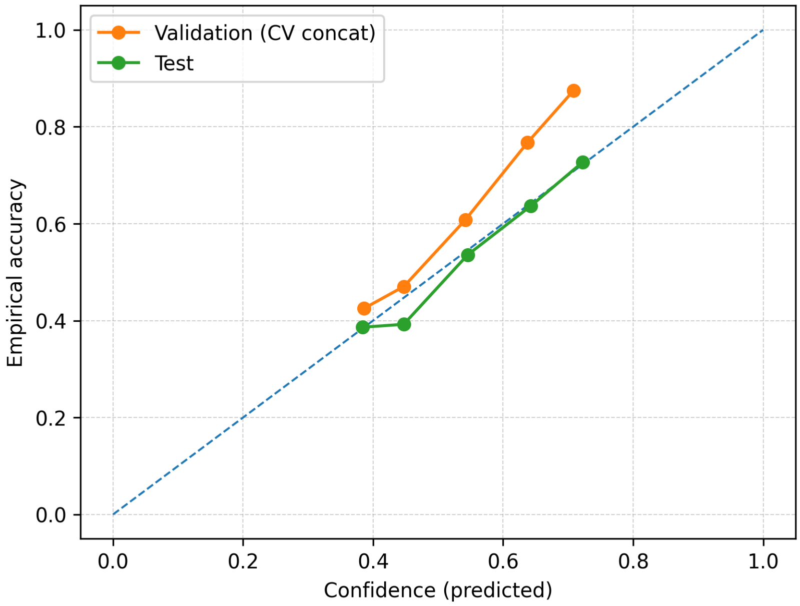


Figure 4.8: Reliability diagram comparing Validation (CV concat) and Test calibration for the Linear SVM SBERT (MPNet) model.

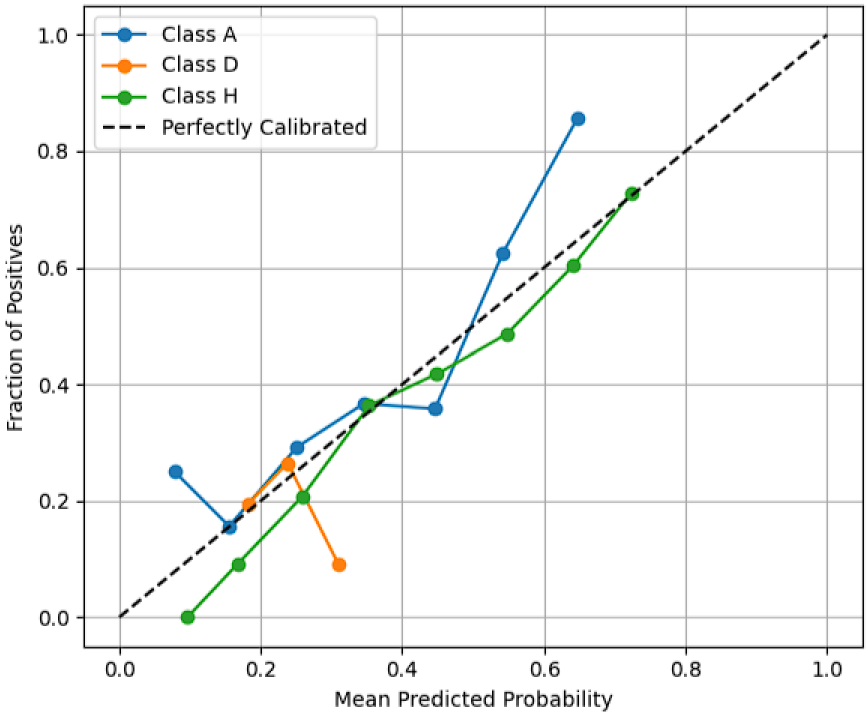


Figure 4.9: Calibration curves of predicted probabilities for the Linear SVM SBERT (MPNet) model.

Table 4.5: Validation, OOF and Test metrics with corresponding deltas for the Linear SVM SBERT (MPNet) model.

Metric	Val (mean ± sd)	OOF	Test	Δ(Test–Val)	Δ(Test–OOF)
Log Loss	1.0028 ± 0.0313	1.0028	1.0313	+0.0285	+0.0285
RPS	0.1415 ± 0.0055	0.1415	0.1444	+0.0029	+0.0029
ECE	0.4804 ± 0.0158	0.4804	0.5050	+0.0246	+0.0246

Table 4.6: Performance of the Linear SVM SBERT (MPNet) model on the holdout 2024/25 season.

Model	Log Loss	ECE	Brier	Bal. Acc.	Macro-F1	Acc.	RPS
Linear SVM	1.0313	0.5050	0.2059	0.4150	0.3558	0.4789	0.1444

class predictions were obtained from the argmax of the predicted probabilities, aligned with the class order (A, D, H). This adjustment ensured that the probability outputs from the statistical and textual models were directly comparable and thus suitable for ensemble and fusion strategies.

The Tables 4.7 and 4.8 present the performance of the hybrid models across the full set of evaluation metrics. Within the ensemble approaches, Soft Voting achieved a Log Loss of 0.9066 and an Accuracy of 58.16%, while Class-Specific Voting delivered very similar results, with a Log Loss of 0.9086 and an Accuracy of 57.89%. In contrast, Dynamic Voting performed weakest in this group, reaching a Log Loss of 0.9259, an Accuracy of 57.63%, and a Macro-F1 of 0.4483.

One of the strongest ensemble results came from stacking with residual learning, which delivered one of the strongest ensemble results, reaching a Log Loss of 0.9043 and an Accuracy of 57.89%. The model also achieved a Balanced Accuracy of 51.39% and a Macro-F1 score of 0.4466. In terms of Log Loss, this approach ranked just behind Early Fusion, indicating that the residual correction strategy offered a measurable advantage over standard voting.

The Early Fusion configuration, in which tweet-derived features were concatenated directly with structured statistical features before training a Logistic Regression model, delivered the strongest overall performance. It achieved the lowest Log Loss (0.8954), the highest Accuracy (59.74%), and the highest Balanced Accuracy (53.57%). Moreover, Macro-F1 increased to 0.4914, surpassing the levels observed in the other hybrid methods. In terms of secondary metrics, Early Fusion reinforced this advantage by recording the lowest Brier Score (0.1758), the highest recall (53.57%), and the highest ROC-AUC Macro (0.7382), while maintaining competitive calibration as reflected by its ECE of 0.5762 and RPS of 0.1171.

In summary, the results show that across the different hybrid approaches the models achieved Accuracy values between 57.63% and 59.74% and Log Loss values between 0.8954 and 0.9259. Notably, Early Fusion consistently delivered the strongest results, with Stacking using residual learning emerging as the next-best configuration. The remaining ensemble methods, namely Soft Voting, Class-Specific Voting, and Dynamic Voting, produced performance clustered within a narrower range, with only moderate differences across metrics.

Table 4.7: Hybrid models – main metrics (holdout 2024/25).

Model	Acc	BalAcc	LogLoss	Macro-F1
Soft Voting	58.16 %	51.64 %	0.9066	0.4537
Class-Specific Voting	57.89 %	51.28 %	0.9086	0.4450
Dynamic Voting	57.63 %	51.09 %	0.9259	0.4483
Stacking	57.89 %	51.39 %	0.9043	0.4466
Early Fusion (LogReg, ws=3)	**59.74 %**	**53.57 %**	**0.8954**	**0.4914**

Table 4.8: Hybrid models – secondary metrics (holdout 2024/25).

Model	Brier	ECE	Prec	Rec	AUC	RPS
Soft Voting	0.1784	0.5804	0.5553	0.5164	0.7254	0.1179
Class-Specific Voting	0.1789	0.5755	0.7188	0.5128	0.7269	0.1181
Dynamic Voting	0.1816	0.5448	0.7154	0.5109	0.7226	0.1209
Stacking	0.1783	0.5897	0.4999	0.5139	0.7262	0.1180
Early Fusion (LogReg, ws=3)	**0.1758**	0.5762	0.5922	**0.5357**	**0.7382**	**0.1171**

4.5 Discussion

One useful way to interpret the evidence is by first considering the simplest reference points. A uniform three-class predictor, together with the empirical class prior, provides clear baselines for Log Loss. Every trained family of models in this study performed well above those benchmarks, indicating that they are indeed capturing meaningful structure in the data rather than relying on class frequencies or random variation. This perspective is important to keep in mind, as it shows that any gains we later discuss represent genuine improvements over naive heuristics.

Within the statistical family, the results align with what is typically reported in tabular sports prediction. Among the different approaches, those that can capture non-linear relationships, most notably SVM with an RBF kernel and boosting, converted short-horizon match features into more reliable probability estimates. This supports the view that interactions between recent form, venue, and opponent quality are rarely linear, and that kernels or trees can capture these effects effectively once the features have been carefully engineered. Even so, a carefully regularised Logistic Regression remained close behind, suggesting that much of the predictive signal is still accessible through a linear boundary when the features are well prepared. In contrast, k-nearest neighbours and multilayer perceptrons tended to underperform, and this happened for understandable reasons. In particular, distance-based comparisons lose discriminative power in high-dimensional, heterogeneous spaces, while small neural networks often require more data and tuning to surpass strong kernels or tree ensembles. As a result, their additional capacity was not fully exploited in this setting.

Focusing on the tweet-based models, the results highlight more clearly how textual information contributes to prediction. In the first place, sentiment features and TF-IDF variants provide a stable foundation, although their differences are small, indicating that polarity and term frequency capture only part of the underlying signal. A different picture emerges with dense embeddings. In particular, by condensing contextual semantics into vector representations, they enable simple classifiers to uncover systematic patterns of language use more effectively. This explains why SBERT-MPNet emerges as the leading method within

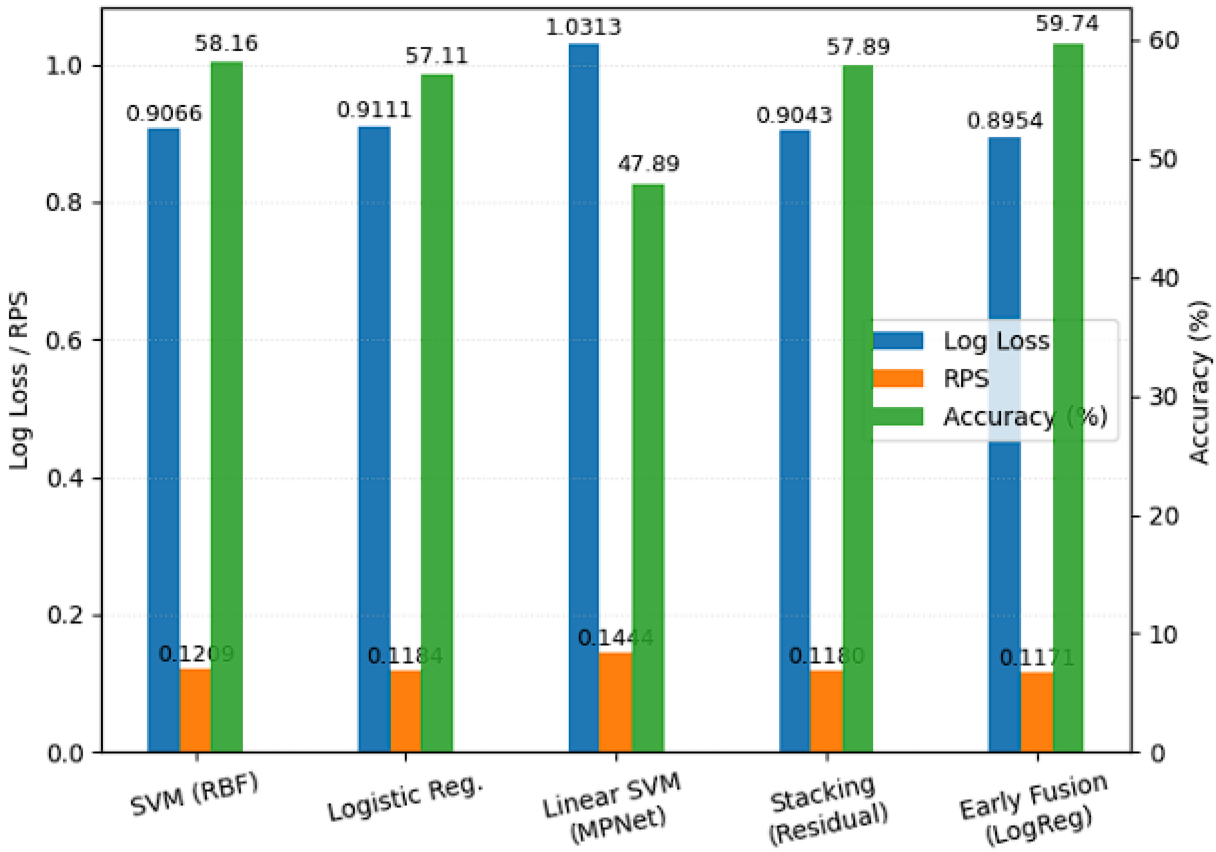


Figure 4.10: Comparison of SVM (RBF), Logistic Regression, Linear SVM (MPNet tweets), Stacking (Residual), and Early Fusion (LogReg) on the 2024/25 holdout: Log Loss and RPS (lower is better) vs. Accuracy.

the textual family, and why replacing Logistic Regression with a Linear SVM further improves performance, since the margin structure in the embedding space is easier to exploit. Equally important, the best tweet-based model behaves consistently across validation, out-of-fold, and test sets, with reliability curves that remain closely aligned between validation and test. In summary, the textual signal generalises in a predictable way, even though its overall strength remains weaker than the best statistical models.

With regard to the hybrid models, although the voting-style hybrids achieved more modest improvements among the ensemble strategies, their behaviour followed a clear and interpretable pattern. In the case of soft voting, class probabilities are averaged across models, and when one component is both more discriminative and better calibrated, the average naturally leans toward it while the tweet model provides only incremental adjustments. When considering class-specific voting, the idea is to assign separate weights for each outcome, which in theory could highlight a class where textual features contribute more strongly, often the draw. In practice, however, because tweet-based probabilities are noisier overall, the learned per-class weights remain conservative, so the ensemble continues to mirror the statistical model in most cases. A different logic applies to dynamic voting, since it adapts weights locally and can respond in situations where the statistical model is uncertain while the tweet signal is more confident. In the present data this mechanism led to selective improvements rather than broad gains, and the ensemble once again tended to rely on the statistical anchor. In summary, the tweet model does not provide enough consistent signal to shift these averaging schemes substantially, so they behave like a cautious statistical framework that only integrates textual cues when the indicators align.

In the case of stacking with residual learning, two independent models are combined: an RBF-SVM trained on structured match statistics and a Linear SVM trained on tweets, with a meta-learner acting as arbiter to decide when to rely on each source of information and when to hold back. Evaluation focuses primarily on Log Loss, since this metric reflects the quality of the full probability distribution rather than only the predicted class. The observed improvements mostly take the form of subtle probability adjustments that reduce Log Loss penalties without requiring a change in the predicted outcome. Actual class flips were rare, and when they did occur they were concentrated in cases where the statistical model was uncertain and the tweet model showed strong confidence. The clearest gains therefore arise when the statistical model's highest and second-highest probabilities are close together, signalling low certainty, while the tweet model provides a decisive signal. On the other hand, when the statistical model is highly confident and the tweets are weak or diffuse, losses appear more frequently, underlining the importance of limiting or even suppressing adjustments in such situations.

When the two base models reach the same conclusion, the impact of stacking is limited. In such cases the ensemble behaves conservatively, preserving the consensus forecast and avoiding unnecessary noise. The main exception is the draw class, where systematic improvements in Log Loss, recall, and Brier suggest that tweets add valuable and complementary information. Prior research has shown that draws are particularly difficult to capture. Baboota and Kaur (2019), for example, found that in Premier League matches precision and recall for draws were drastically lower than for wins: their linear SVM did not predict a single draw, and even the best-performing model, Gradient Boosting, achieved only 0.26 recall. They concluded that models systematically under-predict draws, since these outcomes arise in finely balanced situations that structured statistics alone often fail to represent. Against this background, the fact that stacking improves draw predictions is a notable contribution, as it shows that textual cues, such as references to tight contests, derby tension, or cautious expectations, help to identify signals of balance that the statistical model overlooks.

Even so, not every draw benefited, and in cases where the statistical model was strongly confident against a draw, raising the draw probability could be counterproductive. This highlights the importance of guardrails, ensuring that draw adjustments are made primarily when the statistical margin is narrow and the textual evidence is coherent.

The picture changes when the models disagree. In situations of strong conflict, where both are confident but point to opposite outcomes, the stack tends to improve Log Loss particularly when the tweet model provides the stronger signal. Importantly, the ensemble does not simply follow whichever probability is higher, instead, its behaviour varies across fixtures. In some cases it aligns with tweets, while in others it preserves the statistical forecast, which indicates that the meta-learner has learned more nuanced patterns than a simplistic rule that always favours the most confident source.

A further perspective comes from analysing upsets, where stacking delivers additional gains. Improvements are most evident when tweets align with the non-favourite outcome, plausibly reflecting recent context such as injuries or suspensions that historical tabular features do not capture. Nevertheless, there are risks in atmospheres dominated by collective enthusiasm against clear favourites, where the stack can be nudged toward poorer decisions.

Finally, from a calibration perspective, the strongest improvements appear in cases where the statistical model had been most confident. In these situations, the stack reins in overconfidence by moderating inflated probabilities, which not only sharpens predictions in uncertain conditions but also strengthens probabilistic reliability, as seen in the behaviour of Brier and RPS.

At this stage the focus shifts to an early fusion strategy with VADER. This is not a hybrid ensemble in the strict sense, since the integration takes place before modelling: a single Logistic Regression is trained directly on both the structured match statistics and the prematch sentiment features. The purpose of this setup is not only to test an alternative configuration but also to evaluate more directly the contribution of tweets when incorporated into the feature space from the outset.

Evaluation concentrates on the reduction in overall probability error, measured as the change in Log Loss relative to the statistical baseline. In practice, the early fusion model achieved a consistent drop in error that appears robust rather than coincidental. Most of the improvement came from subtle refinements, as the model adjusted the predicted probabilities in the right direction. It gave slightly more weight to the outcome that eventually materialised and reduced confidence where it had been overstated, while usually preserving the top predicted class. Changes in the leading outcome did occur, but they were relatively infrequent and tended to arise when the statistical model was uncertain while the sentiment provided a clearer signal.

The best results appear when the statistical baseline is undecided, with its two leading options close together, and the sentiment signal points clearly in one direction. In those fixtures, the social signal contributes contextual information that the tables alone cannot capture, such as player absences, morale, or tactical discussions, and the probability distribution improves consistently. By contrast, when the baseline is already very confident and the sentiment is weak or mixed, adjustments rarely add value and can even introduce small degradations, which suggests they should be limited. When both statistical confidence and sentiment are strong and broadly aligned, early fusion behaves more like a fine-tuner, preserving the favourite while improving calibration.

As in the previous stage, draws benefit the most. The baseline tends to under-allocate probability to this outcome, whereas early fusion redistributes a small share of probability mass from wins to draws whenever pre-match discourse points to balance from both sides. This makes the distribution more realistic without overturning the main decision. Another factor to consider is tweet volume, which plays an important role, as low to medium levels often coincide with useful and contextually grounded information that aligns well with situations of low statistical confidence. Very high volumes, however, are more common in high-profile matches and can amplify attention bias or collective enthusiasm, so sentiment should be weighted more cautiously in those situations.

Placing the two hybrid approaches side by side helps to clarify their respective roles. Both deliver improvements mainly through subtle probability adjustments, with positive effects concentrated in situations where the statistical model is less decisive, and both enhance not only discrimination but also calibration, as shown by ECE, Brier, and RPS. In this comparison, stacking with residual learning emerges as a selective calibrator, stepping in under disagreement and reducing overconfidence when needed. Early fusion, by contrast, integrates statistics and sentiment within a single learner and behaves like a contextual fine-tuner, conditioning probability mass on the strength and direction of the social signal. While stacking proves most valuable in strong conflicts between statistical and textual views, early fusion is particularly effective when uncertainty in the statistical model coincides with a clearly asymmetric sentiment landscape.

When considered together with the voting-style ensembles and the tweet-only models, a consistent hierarchy emerges. At the core of this hierarchy are the statistical models, which deliver the most stable predictive power. Complementing them, tweet-based models provide additional context, proving most useful when pre-match narratives are salient or the statistical view is indecisive. The simple voting hybrids largely mirror the statistical component and borrow from text only sparingly, since the textual signal is not uniformly strong. By contrast, stacking and early fusion learn more actively where and how to incorporate that signal, reducing penalties in Log Loss and moderating overconfidence, particularly in draws and in fixtures where information beyond recent numerical form appears to matter, as illustrated in Figure 4.10. Taken together, the evidence indicates that these approaches not only strengthen predictive accuracy but also enhance the reliability of probability estimates, especially in the areas where traditional statistical models are most limited.

Chapter 5

Conclusions

5.1 Introduction

Chapter 5 serves as the conclusion of the study, offering a concise synthesis of what was achieved, the constraints that frame those findings, and the directions that future research might take. Rather than revisiting the detailed analysis from the previous chapter, it summarises the essentials into a clear set of conclusions that link the empirical results to broader questions of football outcome prediction.

The chapter is structured into three sections. The first, Summary and Objectives Achieved, consolidates the main outcomes and clarifies which goals were fully reached and which remain only partially addressed. The second, Limitations, discusses the key factors that constrain interpretation and practical use, including aspects of data availability, modelling choices, and temporal generalisation across seasons. The third, Recommendations for Future Research, outlines practical directions to enhance robustness and extend applicability across leagues and competitive contexts. Finally, the chapter concludes with a section on Security and Ethics, reflecting on the integrity, transparency, and responsible use of the developed models, as well as the broader social implications of data-driven prediction in sports.

Overall, these sections provide a compact closure to the work, while also positioning it within an ongoing research agenda that seeks to improve the reliability, interpretability, and scope of football prediction models.

5.2 Summary and Objectives Achieved

The central research question that guided this study was:

How effective are hybrid models in integrating statistical match data and sentiment analysis for football match outcome predictions compared to traditional methods?

In order to address this question, a set of objectives was established in Chapter 1. These included reviewing prior research, constructing the datasets, implementing sentiment analysis, designing the hybrid predictive models, and evaluating their performance on historical data from the English Premier League (EPL).

The systematic literature review established the state of research on football outcome prediction and revealed a clear gap in the field. While hybrid approaches have been widely explored in other domains, their potential in football prediction has received little attention. This gap provided the main motivation for the project and positioned it as a contribution to an area that remains underexplored.

The project subsequently constructed datasets that combined match statistics with sentiment features aligned at the fixture level. The statistical component captured dimensions such as team form, venue, and opponent quality, while the sentiment indicators were extracted from social media posts and classified as positive, negative, or neutral. Although the initial plan also considered sports news, the analysis ultimately focused on social media, which was processed in full and integrated into the modelling pipeline.

Building on these datasets, a range of machine learning models was developed and tested, covering both linear and non-linear approaches. Logistic regression confirmed its competitiveness in structured feature spaces, whereas methods such as SVM with RBF kernels and boosting ensembles showed greater ability to capture the non-linear relationships present in the data.

To explore their added value, sentiment-based features were integrated with statistical predictors through several hybrid strategies. Among these, early fusion, which concatenates sentiment features with structured inputs, and residual stacking, which adjusts baseline statistical predictions using sentiment information, proved particularly effective. Across evaluations, these hybrid configurations consistently outperformed models trained solely on match statistics.

In terms of evaluation, tests on temporally separated holdout fixtures confirmed that the proposed hybrid models enhanced probabilistic accuracy and, in some cases, classification accuracy. The gains were most evident in matches where the outcome was harder to determine, highlighting the models' ability to capture additional signals under uncertainty. Analyses of feature importance further indicated that sentiment variables contributed meaningfully alongside core statistical predictors, strengthening the case for integrating multiple data sources.

In the context of football outcome prediction, the literature typically reports seasonal out-of-sample benchmarks with Accuracy around 50–55% and Log Loss between 0.95 and 1.05, reflecting both the competitive balance of leagues and the frequency of draws. A widely cited reference is the study by Baboota and Kaur (2019) on the English Premier League, which trained on nine seasons and tested on two complete seasons, obtaining its best performance with Gradient Boosting at a Ranked Probability Score (RPS) of 0.2156 and an accuracy of approximately 58.5%. Against this background, two hybrid configurations stood out in this project. Early fusion with Logistic Regression achieved an Accuracy of 59.74% and a Log Loss of 0.8954, together with an RPS of 0.1171 and a Brier score of 0.1758. Stacking with residual learning also delivered competitive results, confirming its ability to improve probabilistic accuracy and calibration, particularly in matches where the statistical model was less decisive. These results place the hybrid approaches above the typical accuracy range and below the common log-loss interval, providing clear evidence that they deliver highly competitive predictive performance.

In summary, the objectives set out in Chapter 1 were achieved, and the central research question can be answered with clear evidence: hybrid models are effective in football outcome prediction, delivering measurable improvements over approaches based solely on match statistics. Furthermore, the results place the proposed models at a level comparable to, and in some cases exceeding, benchmarks reported in the literature, reinforcing the value of hybrid modelling as a robust strategy for prediction in the English Premier League.

5.3 Limitations

Any predictive approach of this kind carries inherent limitations that need to be acknowledged. Models trained on structured match statistics, for instance, are exposed to shifting dynamics across a season: changes in squads, coaching staff, or fixture congestion can gradually weaken patterns learned from past data. The choice of observation window also introduces trade-offs, since short horizons risk amplifying noise, whereas longer ones weaken the recent performance signal they are meant to preserve. In addition, many statistical features overlap, creating collinearity that complicates interpretation, and class imbalance, particularly the relative rarity of draws, can bias predictions. These models also struggle to incorporate sudden contextual factors such as injuries, tactical adjustments, or rivalry intensity, and they remain vulnerable to information leakage when features are derived too close to betting odds or post-event summaries.

When turning to tweet-based models, a different set of weaknesses becomes apparent. Social media attention is unequally distributed, concentrated on high-profile clubs and matches, while smaller fixtures suffer from sparse data. The medium itself is noisy: sarcasm, slang, bots, and coordinated campaigns can distort sentiment signals, and assigning polarity to the correct team is not always straightforward. Maintaining strict temporal hygiene is especially important, for example, tweets posted a day before the match are valid inputs, but tweets about the confirmed line-up ten minutes before kick-off, or reactions after an early goal, would leak information the model should not have. Without clear cut-offs, results can appear stronger than they would in a genuine prediction scenario. In addition, coverage varies across time, language, and geography, and the population of Twitter users is not representative of football audiences as a whole. These factors make text-derived features fragile and prone to drift as vocabulary and patterns of expression change over time.

Hybrid configurations are designed to balance these weaknesses, yet they are also constrained by them. Simple voting schemes, for example, assume that base models are equally strong and well calibrated, which is rarely the case. Dynamic or class-specific weighting relies on validation patterns that are expected to persist, but under drift these weights can mislead rather than improve. Stacking and early fusion have the potential to deliver calibration gains, although typically under specific conditions and with limited influence on the predicted class itself. Their effectiveness further depends on the calibration of the base models; when one model produces poorly aligned probabilities, its errors tend to propagate through the ensemble.

Another limitation lies in the textual component, which hybrids inevitably inherit. The challenge is not only to combine statistics and tweets but also to ensure that only relevant and informative tweets are included. Irrelevant or noisy content, such as memes, off-topic chatter, or coordinated fan activity, can overwhelm the statistical signal and in some cases lead to worse predictions than statistics alone. More broadly, because most validation is tied to a single league (the EPL) and a single holdout season (2024/25), there remains a constant risk of overfitting to season-specific patterns, which makes generalisation to new competitions or time periods uncertain. The limited data also introduces seed sensitivity and hyperparameter instability, both of which can reduce reproducibility.

Overall, these limitations show that although the models are able to capture meaningful signals, their reliability depends strongly on context, data quality, and temporal robustness.

The findings should therefore be interpreted as evidence of potential rather than as guarantees of stable performance, with careful monitoring and periodic recalibration required before deployment in real-world prediction tasks.

5.4 Recommendations for Future Research

Building on the findings and the limitations outlined above, several promising directions emerge for future research. A first priority lies in expanding and diversifying the data. Since this study focused on the English Premier League and a single holdout season (2024/25), questions naturally arise about temporal validity and cross-league generalisation. Extending the analysis to other competitions and longer time horizons would not only reinforce robustness but also clarify whether the observed patterns are league-specific or hold broader relevance.

Building on this foundation, another line of inquiry concerns textual sources and filtering. While Twitter provided a rich stream of sentiment, it also revealed considerable noise, with results reflecting both its value and its fragility. A natural step forward would be to extend the range of textual evidence by drawing on alternative sources of fan and expert opinion, such as sports news, online forums, or match previews, to complement social media signals. Equally important is the development of more advanced filtering mechanisms applied before the modelling stage. The strategies adopted in this study, including queries based on hashtags, keywords, or mentions, helped reduce noise but were not sufficient to isolate truly informative content. A promising next step would be the design of a reading head or pre-selection module capable of distinguishing relevant tweets from off-topic, sarcastic, or coordinated material, thereby strengthening the signal-to-noise ratio and improving the stability of sentiment features over time.

In parallel, attention should also be directed towards model robustness and calibration. Although hybrid approaches such as stacking and early fusion delivered clear gains, their effectiveness remained highly dependent on the calibration of base models and on context-specific conditions. A valuable direction for future research would be to explore calibration techniques more systematically and to design ensemble combinations capable of adapting dynamically over time. In addition, integrating uncertainty estimates could support the development of more advanced decision-making mechanisms, allowing models to judge when textual signals should be given more weight, reduced in importance, or disregarded according to the scenario. Although the study incorporated feature-importance analyses, a natural next step would be to introduce an explainability layer that links each prediction to its most influential statistical features and tweet signals. These explanations could then be presented through a lightweight language model, capable of generating clear and user-friendly narratives that explain why a given outcome was favoured.

Closely related to this, reproducibility and stability continue to represent important challenges. When data are limited, models may become sensitive to random seeds and hyperparameter choices, which compromises reliability across different runs. Future studies could mitigate these risks by relying on larger datasets, applying repeated validation across multiple seasons, and adopting complementary validation strategies. Such measures would enhance the robustness of findings and increase confidence in the reported improvements.

Overall, these directions define a research agenda that complements the present study, encouraging broader validation across leagues and more sophisticated handling of textual information. Moving forward in this way would not only deepen academic understanding of football outcome prediction but also contribute to the development of more robust and context-aware modelling strategies.

5.5 Security and Ethics

While conducting the study, several ethical and operational considerations were integrated to guarantee the integrity, transparency, and responsible use of the research outcomes. To ensure this, data integrity and authenticity were safeguarded through authenticated data pipelines for both football statistics and social media sources, ensuring that all inputs originated from verified providers. The experimental workflow was designed to be fully reproducible and traceable, with fixed random seeds, version-controlled datasets, and consistent preprocessing stages. These measures enabled every model to be independently replicated and audited, aligning the study with best practices in Responsible AI.

Alongside these safeguards, algorithmic transparency and bias were equally important considerations. The study acknowledged that sentiment analysis tools such as VADER and CardiffNLP can be influenced by linguistic and cultural biases, particularly within the informal and emotionally charged environment of football discussions on social media. To address these challenges, both models were applied and compared to evaluate their ability to capture meaningful sentiment signals. The model that showed greater consistency and sensitivity to contextual sentiment was then selected for subsequent experiments, reflecting a deliberate and transparent effort to mitigate linguistic bias in the analysis.

Beyond considerations of bias and transparency, the ethical framing of the research purpose was equally important. The work was conducted exclusively for academic and educational purposes. Although the implemented models are capable of generating probabilistic forecasts, their primary objective was to explore how match statistics and social sentiment data could be integrated and compared, thereby advancing the understanding of predictive modelling in complex human domains. The research was not intended for commercial or betting-related applications.

If systems of this nature were ever to be deployed publicly, appropriate safeguards would be essential to prevent manipulation or misuse. Such measures would include strict access control and authentication mechanisms, adversarial robustness testing to detect data tampering, anomaly detection systems to flag irregular usage patterns, audit logging to ensure full traceability of predictions, and explainability layers to help users interpret probabilistic outputs. Taken together, these safeguards would reinforce both the security and transparency of AI-driven systems.

At a broader level, social implications also warrant attention, as the misuse of sports prediction tools, although unrelated to gambling, could unintentionally reinforce behaviours associated with betting addiction. To mitigate these risks, responsible deployment practices should be adopted, including clear disclaimers about model uncertainty, limits on access frequency to discourage compulsive use, monitoring of repetitive query patterns that may signal dependency, and user education regarding the experimental and uncertain nature of the forecasts.

By situating the research within a strictly academic context and emphasising the inherent unpredictability of football, this work promotes ethical and socially responsible applications of AI. Even with advanced predictive tools, the dynamics of sports betting remains highly complex and uncertain, underscoring the limits of data-driven approaches in this domain.

Bibliography

Abdullah, A Sheik et al. (Aug. 2024). “Enhancing Predictive Analytics Through Hybrid Machine Learning Models: A Comparative Analysis Of SVM-NB And Lasso-Ridge Techniques”. In: *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)*. Bengaluru, India: IEEE, pp. 1–10. isbn: 979-8-3503-7289-2. doi: 10.1109/NMITCON62075.2024.10699192. (Visited on 01/25/2025).

AlMulla, Jassim et al. (Aug. 2023). “SoccerNet: A Gated Recurrent Unit-based Model to Predict Soccer Match Winners”. In: *PLOS ONE* 18.8. Ed. by Rahul Gomes, e0288933. issn: 1932-6203. doi: 10.1371/journal.pone.0288933. (Visited on 01/03/2025).

Angelini, Giovanni and Luca De Angelis (2016). *PARX Model for Football Matches Predictions*. doi: 10.13140/RG.2.2.27404.72325. (Visited on 01/25/2025).

Baboota, Rahul and Harleen Kaur (Apr. 2019). “Predictive Analysis and Modelling Football Results Using Machine Learning Approach for English Premier League”. In: *International Journal of Forecasting* 35.2, pp. 741–755. issn: 01692070. doi: 10.1016/j.ijforecast.2018.01.003. (Visited on 01/03/2025).

Baratela, Eduardo Alves et al. (Sept. 2024). *Predicting Soccer Matches with Complex Networks and Machine Learning*. doi: 10.48550/arXiv.2409.13098. arXiv: 2409.13098 [cs]. (Visited on 01/03/2025).

Beal, Ryan et al. (Dec. 2020). *Combining Machine Learning and Human Experts to Predict Match Outcomes in Football: A Baseline Model*. doi: 10.48550/arXiv.2012.04380. arXiv: 2012.04380 [cs]. (Visited on 01/03/2025).

Berrar, Daniel, Philippe Lopes, and Werner Dubitzky (Jan. 2019). “Incorporating Domain Knowledge in Machine Learning for Soccer Outcome Prediction”. In: *Machine Learning* 108.1, pp. 97–126. issn: 0885-6125, 1573-0565. doi: 10.1007/s10994-018-5747-8. (Visited on 01/03/2025).

Byeon, Haewon (Oct. 2021). “Development of a Predictive Model for Mild Cognitive Impairment in Parkinson’s Disease with Normal Cognition Using Kernel-Based C5.0 Machine Learning Blending: Preliminary Research”. In: *The 2nd International Electronic Conference on Applied Sciences*. MDPI, p. 18. doi: 10.3390/ASEC2021-11147. (Visited on 01/25/2025).

Capobianco, Giovanni et al. (2019). “Can Machine Learning Predict Soccer Match Results?” In: *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*. Prague, Czech Republic: SCITEPRESS - Science and Technology Publications, pp. 458–465. isbn: 978-989-758-350-6. doi: 10.5220/0007307504580465. (Visited on 01/03/2025).

Chen, Hao-Yuan et al. (Dec. 2023). *Hybrid Quantum Neural Network in High-dimensional Data Classification*. doi: 10.48550/arXiv.2312.01024. arXiv: 2312.01024 [cs]. (Visited on 01/25/2025).

Chen, T. et al. (Dec. 2019). *Hybrid Machine Learning Models of Classifying Residential Requests for Smart Dispatching*. doi: 10.48550/arXiv.1912.10546. arXiv: 1912.10546 [cs]. (Visited on 01/25/2025).

Constantinou, Anthony C. (Jan. 2019). "Dolores: A Model That Predicts Football Match Outcomes from All over the World". In: *Machine Learning* 108.1, pp. 49–75. issn: 0885-6125, 1573-0565. doi: 10.1007/s10994-018-5703-7. (Visited on 01/03/2025).

Da Costa, Daniel C., Ricardo Prudêncio, and Alexandre Mota (Dec. 2023). "Assessor Models with a Reject Option for Soccer Result Prediction". In: *2023 International Conference on Machine Learning and Applications (ICMLA)*. Jacksonville, FL, USA: IEEE, pp. 1200–1205. isbn: 979-8-3503-4534-6. doi: 10.1109/ICMLA58977.2023.00180. (Visited on 01/03/2025).

Dahiya, Shashi, S. S Handa, and N. P. Singh (Oct. 2015). "Credit Modelling Using Hybrid Machine Learning Technique". In: *2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI)*. Faridabad, India: IEEE, pp. 103–106. isbn: 978-1-4673-6792-9. doi: 10.1109/ICSCTI.2015.7489612. (Visited on 01/25/2025).

Dip, Amartya Das, Nasimur Rahman, and Mohiuddin Ahmed (Sept. 2024). "Predicting Football Match Results: An Analysis of Feature Selection and Machine Learning Techniques Using a Curated Dataset". In: *2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON)*. Rajshahi, Bangladesh: IEEE, pp. 927–932. isbn: 979-8-3315-1798-4. doi: 10.1109/PEEIACON63629.2024.10800599. (Visited on 01/03/2025).

Elmiligi, Haytham and Sherif Saad (Jan. 2022). "Predicting the Outcome of Soccer Matches Using Machine Learning and Statistical Analysis". In: *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*. Las Vegas, NV, USA: IEEE, pp. 1–8. isbn: 978-1-6654-8303-2. doi: 10.1109/CCWC54503.2022.9720896. (Visited on 01/03/2025).

Giroh, Himanshu, Vipin Kumar, and Gurdiyal Singh (Aug. 2023). "Improving the Performance of Hybrid Models Using Machine Learning and Optimization Techniques". In: *International Journal of Membrane Science and Technology* 10.2, pp. 3396–3409. issn: 2410-1869. doi: 10.15379/ijmst.v10i2.3138. (Visited on 01/25/2025).

Gorczyca, Michael T., Nicole C. Toscano, and Julius D. Cheng (May 2019). "The Trauma Severity Model: An Ensemble Machine Learning Approach to Risk Prediction". In: *Computers in Biology and Medicine* 108, pp. 9–19. issn: 00104825. doi: 10.1016/j.compbiomed.2019.02.025. (Visited on 01/25/2025).

Goto, Shingo and Toru Yamada (July 2023). "What Drives Biased Odds in Sports Betting Markets: Bettors' Irrationality and the Role of Bookmakers". In: *International Review of Economics & Finance* 86, pp. 252–270. issn: 10590560. doi: 10.1016/j.iref.2023.03.002. (Visited on 09/11/2025).

Groll, Andreas et al. (Oct. 2019). "A Hybrid Random Forest to Predict Soccer Matches in International Tournaments". In: *Journal of Quantitative Analysis in Sports* 15.4, pp. 271–287. issn: 1559-0410, 2194-6388. doi: 10.1515/jqas-2018-0060. (Visited on 01/26/2025).

Gudla, Sirisha et al. (Aug. 2023). "Leveraging Machine Learning Algorithms for Football Predictions and Wager Suggestions". In: *2023 9th International Conference on Smart Computing and Communications (ICSCC)*. Kochi, Kerala, India: IEEE, pp. 50–54. isbn: 979-8-3503-1409-0. doi: 10.1109/ICSCC59169.2023.10335054. (Visited on 01/03/2025).

Hassard, Peter and Dermot Kerr (June 2024). "Predicting Football Match Outcomes Using Event Data and Machine Learning Algorithms". In: *2024 35th Irish Signals and Systems Conference (ISSC)*. Belfast, United Kingdom: IEEE, pp. 1–6. isbn: 979-8-3503-5298-6. doi: 10.1109/ISSC61953.2024.10603147. (Visited on 01/03/2025).

He, Bohao et al. (Feb. 2023). "Mapping Seagrass Habitats of Potential Suitability Using a Hybrid Machine Learning Model". In: *Frontiers in Ecology and Evolution* 11, p. 1116083. issn: 2296-701X. doi: 10.3389/fevo.2023.1116083. (Visited on 01/25/2025).

Hegarty, Tadgh and Karl Whelan (Apr. 2025). "Forecasting Soccer Matches with Betting Odds: A Tale of Two Markets". In: *International Journal of Forecasting* 41.2, pp. 803–820. issn: 01692070. doi: 10.1016/j.ijforecast.2024.06.013. (Visited on 09/11/2025).

Hu, Sicheng and Min Fu (Aug. 2022). "Football Match Results Predicting by Machine Learning Techniques". In: *2022 International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI)*. Zakopane, Poland: IEEE, pp. 72–76. isbn: 978-1-6654-5470-4. doi: 10.1109/ICDACAI57211.2022.00022. (Visited on 01/03/2025).

Iyer, Varun (Mar. 2024). *A Comparative Analysis of Sentiment Classification Models for Improved Performance Optimization*. doi: 10.36227/techrxiv.171073040.03879551/v1. (Visited on 01/25/2025).

Iyiola, Tomilayo P., Hilary I. Okagbue, and Oluwole A. Odetunmibi (Nov. 2022). "USE OF THE FIRST AND SECOND HALVES RESULTS TO CLASSIFY THE FINAL OUTCOME OF ENGLISH PREMIER LEAGUE MATCHES". In: *Advances and Applications in Statistics* 82, pp. 53–64. issn: 09723617. doi: 10.17654/0972361722080. (Visited on 01/03/2025).

Kazienko, Przemyslaw, Edwin Lughofer, and Bogdan Trawinski (Dec. 2015). "Editorial on the Special Issue "Hybrid and Ensemble Techniques in Soft Computing: Recent Advances and Emerging Trends"". In: *Soft Computing* 19.12, pp. 3353–3355. issn: 1432-7643, 1433-7479. doi: 10.1007/s00500-015-1916-x. (Visited on 01/25/2025).

Kinalioğlu, İsmail and Coskun Kuş (Jan. 2023). "Prediction of Football Match Results by Using Artificial Intelligence-Based Methods and Proposal of Hybrid Methods". In: *International Journal of Nonlinear Analysis and Applications* 14.1. doi: 10.22075/ijnaa.2023.26802.3413. (Visited on 01/03/2025).

Li, Jinchi, Eric Zhao, and Zhixiang Li (Oct. 2024). *PREDICTING FOOTBALL MATCH OUTCOMES USING LARGE LANGUAGE MODELS: A COMPARATIVE STUDY WITH TRADITIONAL MACHINE LEARNING METHODS*. doi: 10.31235/osf.io/e5wpy. (Visited on 01/25/2025).

Macrì Demartino, Roberto, Leonardo Egidi, and Nicola Torelli (Dec. 2024). "Alternative Ranking Measures to Predict International Football Results". In: *Computational Statistics*. issn: 0943-4062, 1613-9658. doi: 10.1007/s00180-024-01585-z. (Visited on 01/03/2025).

Madeswaran, Tamilselvi et al. (Oct. 2023). "Breast Cancer Prediction Using Hybrid Machine Learning and Nature-Inspired Adaptive Optimization Algorithms". In: *Journal of Chemical Health Risks*. doi: 10.53555/jchr.v13.i4s.1612. (Visited on 01/25/2025).

Malamatinos, Marios-Christos, Eleni Vrochidou, and George A. Papakostas (Aug. 2022). "On Predicting Soccer Outcomes in the Greek League Using Machine Learning". In: *Computers* 11.9, p. 133. issn: 2073-431X. doi: 10.3390/computers11090133. (Visited on 01/03/2025).

Miranda-Peña, Clarissa et al. (2021). "Predicting Soccer Results Through Sentiment Analysis: A Graph Theory Approach". In: *Computational Science – ICCS 2021*. Ed. by Maciej Paszynski et al. Vol. 12747. Cham: Springer International Publishing, pp. 422–435. isbn: 978-3-030-77979-5 978-3-030-77980-1. doi: 10.1007/978-3-030-77980-1_32. (Visited on 01/03/2025).

Moher, David et al. (July 2009). "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement". In: *PLoS Medicine* 6.7, e1000097. issn: 1549-1676. doi: 10.1371/journal.pmed.1000097. (Visited on 01/25/2025).

Mosavi, Amir et al. (Apr. 2019). "State of the Art of Machine Learning Models in Energy Systems, a Systematic Review". In: *Energies* 12.7, p. 1301. issn: 1996-1073. doi: 10.3390/en12071301. (Visited on 01/25/2025).

Narsis, Ouassila Labbani, Erik Dujardin, and Christophe Nicolle (Dec. 2023). “Objective-Driven Modular and Hybrid Approach Combining Machine Learning and Ontology”. In: *2023 15th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)*. Bali, Indonesia: IEEE, pp. 300–305. isbn: 979-8-3503-8382-9. doi: 10.1109/IIAI-AAI-Winter61682.2023.00062. (Visited on 01/25/2025).

Nazim Razali, Muhammad et al. (2022). “Football Matches Outcomes Prediction Based on Gradient Boosting Algorithms and Football Rating System”. In: *13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022)*. doi: 10.54941/ahfe1002524. (Visited on 01/26/2025).

Nimma, Divya and Arjun Uddagiri (Oct. 2024). “Enhancing Tackle Prediction in NFL Games Through Feature Engineering and Hybrid Machine Learning Models”. In: *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. Kirtipur, Nepal: IEEE, pp. 1913–1919. isbn: 979-8-3503-7642-5. doi: 10.1109/I-SMAC61858.2024.10714680. (Visited on 01/25/2025).

Nivetha, S.K. et al. (Jan. 2022). “A Deep Learning Framework for Football Match Prediction”. In: *2022 International Conference on Computer Communication and Informatics (ICCCI)*. Coimbatore, India: IEEE, pp. 1–7. isbn: 978-1-6654-8035-2. doi: 10.1109/ICCCI54379.2022.9740760. (Visited on 01/03/2025).

Nosratabadi, Saeed, Amirhosein Mosavi, et al. (Oct. 2020). “Data Science in Economics: Comprehensive Review of Advanced Machine Learning and Deep Learning Methods”. In: *Mathematics* 8.10, p. 1799. issn: 2227-7390. doi: 10.3390/math8101799. (Visited on 01/25/2025).

Nosratabadi, Saeed, Karoly Szell, et al. (Oct. 2020). “Comparative Analysis of ANN-ICA and ANN-GWO for Crop Yield Prediction”. In: *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*. Ho Chi Minh, Vietnam: IEEE, pp. 1–5. isbn: 978-1-7281-5377-3. doi: 10.1109/RIVF48685.2020.9140786. (Visited on 01/25/2025).

P, Athish V, Rajeswari D, and Sree Nandha S S (Mar. 2023). “Football Prediction System Using Gaussian Naïve Bayes Algorithm”. In: *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*. Tuticorin, India: IEEE, pp. 1640–1643. isbn: 979-8-3503-4664-0. doi: 10.1109/ICEARS56392.2023.10085510. (Visited on 01/03/2025).

Pang, Der-Jiun (Mar. 2023). “Hybrid Machine Learning Model Performance in IT Project Cost and Duration Prediction”. In: *Advances in Science, Technology and Engineering Systems Journal* 8.2, pp. 108–115. issn: 24156698, 24156698. doi: 10.25046/aj080212. (Visited on 01/25/2025).

Peffers, Ken et al. (Dec. 2007). “A Design Science Research Methodology for Information Systems Research”. In: *Journal of Management Information Systems* 24.3, pp. 45–77. issn: 0742-1222, 1557-928X. doi: 10.2753/MIS0742-1222240302. (Visited on 01/24/2025).

Peters, George and Diogo Pacheco (Jan. 2023). *Betting the System: Using Lineups to Predict Football Scores*. doi: 10.48550/arXiv.2210.06327. arXiv: 2210.06327 [cs]. (Visited on 01/25/2025).

Rahman, Riaz et al. (June 2024). “Understanding and Predicting Pregnancy Termination in Bangladesh: A Comprehensive Analysis Using a Hybrid Machine Learning Approach”. In: *Medicine* 103.26, e38709. issn: 0025-7974, 1536-5964. doi: 10.1097/MD.0000000000038709. (Visited on 01/25/2025).

Raju, Muntaqim Ahmed et al. (Dec. 2020). “Predicting the Outcome of English Premier League Matches Using Machine Learning”. In: *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*. Dhaka, Bangladesh: IEEE, pp. 1–6. isbn: 978-1-6654-0489-1. doi: 10.1109/STI50764.2020.9350327. (Visited on 01/03/2025).

Razali, Nazim et al. (2022). “Deep Learning for Football Outcomes Prediction Based on Football Rating System”. In: *10TH INTERNATIONAL CONFERENCE ON APPLIED SCIENCE AND TECHNOLOGY*. Kuala Lumpur, Malaysia, p. 040007. doi: 10.1063/5.0104587. (Visited on 01/03/2025).

Ren, Yiming and Teo Susnjak (Nov. 2022). *Predicting Football Match Outcomes with eXplainable Machine Learning and the Kelly Index*. doi: 10.48550/arXiv.2211.15734. arXiv: 2211.15734 [cs]. (Visited on 01/03/2025).

Rodrigues, Fátima and Ângelo Pinto (2022). “Prediction of Football Match Results with Machine Learning”. In: *Procedia Computer Science* 204, pp. 463–470. issn: 18770509. doi: 10.1016/j.procs.2022.08.057. (Visited on 01/03/2025).

Rose, J. Dafni et al. (July 2022). “Comparison of Football Results Using Machine Learning Algorithms”. In: *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*. Chennai, India: IEEE, pp. 1–7. isbn: 978-1-6654-7413-9. doi: 10.1109/ICSES55317.2022.9914265. (Visited on 01/03/2025).

Satyapanich, Taneeya and Autsadin Somkheawwan (Sept. 2023). “Predicting Game Results for Football League Using Deep Learning”. In: *2023 27th International Computer Science and Engineering Conference (ICSEC)*. Samui Island, Thailand: IEEE, pp. 1–6. isbn: 979-8-3503-4210-9. doi: 10.1109/ICSEC59635.2023.10329770. (Visited on 01/03/2025).

Sreenivasgoud, Pulluri et al. (2024). “Analyzing and Predicting Football Match Results Using Deep Learning”. In: *7TH INTERNATIONAL CONFERENCE ON NANOSCIENCE AND NANOTECHNOLOGY*. Chennai, India, p. 020042. doi: 10.1063/5.0195964. (Visited on 01/03/2025).

Tammouch, Ilyas, Abdelamine Elouafi, and Ibtissam Essadik (Dec. 2024). “Betting on Machine Learning: Extracting Patterns from Football’s Anarchic Odds”. In: *2024 7th International Conference on Advanced Communication Technologies and Networking (CommNet)*. Rabat, Morocco: IEEE, pp. 1–7. isbn: 979-8-3503-6702-7. doi: 10.1109/CommNet63022.2024.10793344. (Visited on 01/03/2025).

Taşpinar, Yavuz Selim, İlkay Çinar, and Murat Koklu (June 2021). “Improvement of Football Match Score Prediction by Selecting Effective Features for Italy Serie A League”. In: *MANAS Journal of Engineering* 9.1, pp. 1–9. issn: 1694-7398. doi: 10.51354/mjen.802818. (Visited on 01/03/2025).

Verdonck, Tim et al. (July 2024). “Special Issue on Feature Engineering Editorial”. In: *Machine Learning* 113.7, pp. 3917–3928. issn: 0885-6125, 1573-0565. doi: 10.1007/s10994-021-06042-2. (Visited on 01/25/2025).

Wang, Fuhai (Nov. 2024). “Comparative Evaluation of Sentiment Analysis Methods: From Traditional Techniques to Advanced Deep Learning Models”. In: *Applied and Computational Engineering* 105.1, pp. 23–29. issn: 2755-2721, 2755-273X. doi: 10.54254/2755-2721/105/2024TJ0056. (Visited on 01/25/2025).

Wang, Tianyou, Zheng Zhang, and Shengxin Zhu (Oct. 2024). “Machine Learning-Based Football Match Prediction System”. In: *Applied and Computational Engineering* 92.1, pp. 181–186. issn: 2755-2721, 2755-273X. doi: 10.54254/2755-2721/92/20241749. (Visited on 01/25/2025).

Wang, Zhe et al. (Oct. 2023). *TacticAI: An AI Assistant for Football Tactics*. doi: 10.48550/arXiv.2310.10553. arXiv: 2310.10553 [cs]. (Visited on 01/25/2025).

Wunderlich, Fabian and Daniel Memmert (Dec. 2022). "A Big Data Analysis of Twitter Data during Premier League Matches: Do Tweets Contain Information Valuable for in-Play Forecasting of Goals in Football?" In: *Social Network Analysis and Mining* 12.1, p. 23. issn: 1869-5450, 1869-5469. doi: `10.1007/s13278-021-00842-z`. (Visited on 01/03/2025).

Yeung, Calvin C. K., Rory Bunker, and Keisuke Fujii (Apr. 2023). "A Framework of Interpretable Match Results Prediction in Football with FIFA Ratings and Team Formation". In: *PLOS ONE* 18.4. Ed. by Anwar P.P. Abdul Majeed, e0284318. issn: 1932-6203. doi: `10.1371/journal.pone.0284318`. (Visited on 01/03/2025).

Zheng, Jiacheng (Nov. 2023). "Soccer Match Prediction Based on Big Data Mining and Multimodal Features". In: *2023 3rd International Signal Processing, Communications and Engineering Management Conference (ISPCEM)*. Montreal, QC, Canada: IEEE, pp. 36–40. isbn: 979-8-3503-0702-3. doi: `10.1109/ISPCEM60569.2023.00013`. (Visited on 01/03/2025).