FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Evaluating the Impact of LLM-Derived Contextual Features from Media Content on Football Player Performance Prediction

Tomás Eiras Silva Martins

Mestrado em Engenharia Informática e Computação

Supervisor: Prof. André Restivo

July 23, 2026

Evaluating the Impact of LLM-Derived Contextual Features from Media Content on Football Player Performance Prediction

Tomás Eiras Silva Martins

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Sérgio Nunes

Examiner: Prof. Fátima Rodrigues

Member: Prof. André Restivo

July 23, 2026

Resumo

A previsão do desempenho de jogadores no futebol assenta quase exclusivamente em estatísticas estruturadas e baseadas em contagens, que registam o que um jogador fez mas não codificam nenhuma das condições externas que moldam o desempenho, tais como a fadiga, as mudanças de função, a confiança e a importância do jogo, aspetos discutidos na cobertura editorial mas ausentes dos conjuntos de dados estruturados. Esta dissertação avalia se sinais contextuais extraídos de notícias pré-jogo por um modelo de linguagem de grande escala (LLM) reduzem o erro de previsão da nota do jogo seguinte quando combinados com estatísticas convencionais. Recorrendo a dados da Liga Portugal da época 2024/2025 e às primeiras vinte e sete jornadas de 2025/2026, um framework de aumento de atributos solicita ao modelo a produção de um vetor contextual de sete dimensões, e três variantes isolam o contributo da recodificação estatística, das evidências em bruto dos artigos e das evidências dos artigos verificadas estatisticamente. Modelos específicos por posição, que combinam um ramo recorrente para o histórico de jogos com um ramo feedforward para o contexto pré-jogo, preveem a nota do jogo seguinte. Uma auditoria de fiabilidade revelou uma exatidão direcional de 80.0% e uma pontuação média por caso de 0.84 em 1.0, confirmando que a extração é coerente, embora irregular entre os construtos. O benefício preditivo revelou-se condicional: os atributos com referência cruzada reduziram o RMSE dos médios em 3.4% face à linha de base estatística e em 4.3% face à linha de referência da média, reduções que um teste de permutação de Monte Carlo confirma como estatisticamente significativas, enquanto os avançados não apresentaram qualquer ganho com a referência cruzada e os defesas melhoraram apenas com a recodificação. Em comparação com a linha de referência que utiliza a média das notas dos últimos cinco jogos de cada jogador, nenhuma configuração neural superou essa referência para os defesas ou os avançados. A melhor variante para os avançados ficou 11.1% acima, um défice estatisticamente significativo, enquanto as melhores variantes para os defesas ficaram 2.6% acima, uma margem estatisticamente indistinguível da própria referência. O trabalho estabelece que o contexto mediático derivado de LLM é recuperável e globalmente positivo para a previsão da nota do jogo seguinte apenas quando as evidências editoriais são verificadas face ao registo estatístico, e apenas para as funções cujas notas refletem o envolvimento individual contínuo.

Palavras-chave: Análise de Dados no Futebol • Previsão do Desempenho de Jogadores • Modelos de Linguagem de Grande Escala (LLMs) • Extração de Atributos Contextuais • Dados Não Estruturados • Fusão Multimodal de Atributos

Abstract

Player performance prediction in football relies almost exclusively on structured, count-based statistics that record what a player did but encode none of the external conditions shaping performance, such as fatigue, role changes, confidence, and fixture stakes, which are discussed in editorial reporting yet absent from structured datasets. This dissertation evaluates whether contextual signals extracted from pre-match news articles by a large language model (LLM) reduce next-match rating prediction error when combined with conventional statistics. Using Liga Portugal data from the 2024/2025 season and the first twenty-seven matchdays of 2025/2026, a feature augmentation framework prompts the model to produce a seven-dimensional contextual vector, and three variants isolate the contribution of statistical re-encoding, raw article evidence, and statistically verified article evidence. Position-specific models combining a recurrent branch for match history with a feed-forward branch for pre-match context forecast the following fixture's rating. A reliability audit found 80.0% directional accuracy and a mean per-case score of 0.84 out of 1.0, confirming that extraction is coherent though uneven across constructs. The predictive benefit proved conditional: cross-referenced features reduced midfielder RMSE by 3.4% against the Statistical Baseline and 4.3% against the Average Baseline, which Monte Carlo permutation test confirms as statistically significant, while forwards showed no gain from cross-referencing and defenders improved only from re-encoding. Against the stronger Average Baseline, no neural configuration surpassed that benchmark for defenders or forwards: the best forward variant remained 11.1% above it, a statistically significant deficit, while the best defender variants trailed by 2.6%, a margin statistically indistinguishable from the benchmark itself. The work establishes that LLM-derived media context is recoverable and net positive for next-match rating prediction only where editorial evidence is verified against the statistical record, and only for roles whose ratings reflect continuous individual involvement.

Keywords: Football Analytics • Player Performance Prediction • Large Language Models (LLMs) • Contextual Feature Extraction • Unstructured Data • Multimodal Feature Fusion

The United Nations Sustainable Development Goals

The United Nations Sustainable Development Goals (SDG) provide a global framework for achieving a better and more sustainable future for all. It includes 17 goals (https://sdgs.un.org/goals) to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice.

This work contributes to the broader objectives of two SDGs. The development of data-driven performance prediction tools in football touches on SDG 8, given the sport's role as a major economic sector and employer. At the same time, the application of open institutional data infrastructure and AI to improve analytical capabilities in sports aligns with SDG 9, which promotes inclusive and sustainable industrialisation and fosters innovation.

The specific Sustainable Development Goals mentioned have the following names:

SDG 8 Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialisation and foster innovation

SDG	Target	Contribution	Performance Indicators and Metrics
8	8.2	Predictive modelling of player performance supports more evidence-based decision-making in player scouting, contract valuation, and squad management, contributing to productivity improvements across the football industry.	Adoption rate of data-driven scouting tools in professional clubs; reduction in scouting errors as measured against transfer outcomes.
9	9.5	The integration of LLM-derived contextual features with structured match data advances applied AI research in sports analytics, demonstrating a replicable framework for combining unstructured and structured data sources in predictive systems.	Reduction in next-match rating prediction error (RMSE) relative to statistical baseline; reproducibility of the pipeline across seasons and competitions.

Acknowledgements

I would first like to thank my supervisor, Professor André Restivo, for all his help and constant availability throughout this dissertation. His guidance at every stage was central to bringing this work to completion.

I am also grateful to zerozero, and especially to Marco and Pedro, for the opportunity to carry out this thesis, and to the rest of the team for their help throughout its development.

To my friends António, Dani, Cardoso, and Zé, thank you for all the moments our friendship brought over the course of the degree, and for the help you gave me whenever I needed it. These years would not have been the same without you.

To my parents and my two twin sisters, Maria and Marta, thank you for believing in me and for supporting me throughout my life, and the same goes to the rest of my close family. I would not have reached this point without you behind me.

Finally, to my girlfriend Rita, thank you for being by my side and for cheering me on in both the happy and the difficult moments. Having you beside me made the hard times lighter and the good ones better.

Tomás Martins

Declaração de Honra

Declaro que a presente dissertação é de minha autoria e não foi previamente apresentada e avaliada nesta ou em outra instituição. As referências a outros autores (afirmações, ideias, pensamentos), assim como a utilização de trabalhos publicados respeitam escrupulosamente as regras da atribuição, e encontram-se devidamente indicadas no texto e nas referências bibliográficas, de acordo com as normas de referenciação. Tenho consciência de que a prática de plágio ou de autoplágio constituem ilícitos académicos, conforme estabelecido no Código de Ética da U.Porto.

Declaro, ainda, que utilizei ferramentas de Inteligência Artificial para a realização de parte(s) da presente dissertação, e essa utilização e os seus resultados foram objeto de cuidada verificação para deteção de erros, imprecisões, atribuições infundadas ou outras formas de enviesamento de conteúdos e argumentos.

Tomás Eiras Silva Martins

"The ball is round, the game lasts ninety minutes, and everything else is just theory"

Sepp Herberger

Contents

List of Figures

List of Tables

List of Acronyms

1D-CNN	One-Dimensional Convolutional Neural Network
ABPNN	Adaptive Back Propagation Neural Network
aGg	Actual Goals per game
AI	Artificial Intelligence
CSL	Chinese Super League
DF	Defender
FGpct	Field Goal Percentage
FN	False Negative
FP	False Positive
FW	Forward
GCL	Graph Convolutional Layer
GK	Goalkeeper
GNN	Graph Neural Network
GPS	Global Positioning System
LLM	Large Language Model
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MCC	Matthews Correlation Coefficient
MF	Midfielder
ML	Machine Learning
MLP	Multi-Layer Perceptron
MLR	Multiple Linear Regression
MSE	Mean Squared Error
NBA	National Basketball Association
PCA	Principal Component Analysis
PIM	Player Impact Metric
PPR	Personalized PageRank
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF	Random Forest
RMSE	Root Mean Squared Error

RNN	Recurrent Neural Network
SCS	Sequential Consistency Score
SDG	Sustainable Development Goals
SLR	Systematic Literature Review
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UFC	Ultimate Fighting Championship
xG	Expected Goals
xGg	Expected Goals per game
xT	Expected Threat
XGB	Extreme Gradient Boosting (XGBoost)

Chapter 1

Introduction

Football has become one of the most data-rich sports in the world, yet the ability to predict individual player performance remains limited by the kinds of information that current models can consume. Most prediction systems rely on structured match statistics that record observable outputs but carry no encoding of the external circumstances that shape them. This dissertation investigates whether large language models can close that gap by extracting contextual signals from unstructured media and encoding them as additional features alongside conventional statistical data. The following sections frame the context in which this work is situated, the motivation behind it, the problem it addresses, the research questions that guide the evaluation, and the structure of the remainder of the document.

1.1 Context

Football is among the most widely followed sports in the world, drawing billions of casual viewers and committed supporters alike [17], and over recent decades it has grown into an industry whose economic weight rivals that of major entertainment sectors. During the 2023/2024 season Real Madrid became the first club to pass €1 billion in revenue in a single year, and the highest-earning clubs together reached a record €11.2 billion [14], figures that convey the commercial scale at which the sport now operates.

Much of that growth has been shaped by technology, which has turned football into a heavily data-driven sport. The actions of professional players, in training and in competition, are now tracked through Global Positioning System (GPS) devices, inertial sensors, and event-coding systems that record physical load, positional movement, technical execution, and tactical behaviour. The volume and granularity of this data have changed how clubs reach decisions [10, 26, 32], and they have supported the rise of platforms such as Opta [42], zerozero [65], and FBref [16], through which fans, analysts, and researchers examine dynamics of the game that are not visible from the stands alone.

The same abundance of data has reshaped neighbouring industries. Betting operators draw on detailed statistics to price odds and to open markets on finer events such as shots, corners, fouls,

and possession, which lets users base decisions on evidence rather than intuition [46]. Fantasy football rests on the same foundation, since the fairness of those competitions depends on accurate and current player data that drives scoring and team selection. In both cases access to performance data has pushed the activity toward more strategic decisions grounded in the numbers.

1.2 Motivation

Raw data, however large, carries little value until it is interpreted, and the volume now produced in football has long passed the point at which patterns can be recovered by hand. This is where Artificial Intelligence (AI) and Machine Learning (ML) become useful, since algorithms can surface relationships across large datasets that would otherwise stay hidden and turn them into predictive signals that support scouting, tactical planning, and performance evaluation [41], often across spans of matches, seasons, and entire leagues.

The clearest prize is accurate forecasting. Being able to anticipate how a player will develop or perform, or how a fixture will unfold [11], carries direct value for clubs, coaches, and analysts, and advances in predictive modelling could reshape talent identification, opposition analysis, and match preparation [53]. Greater accuracy also tends to attract further investment into the analytical side of the sport.

Predictive analytics may also narrow the financial distance between smaller clubs and the wealthiest ones. Clubs without the means to sign established stars can instead use data to find undervalued players and to judge their likely development, an approach already followed by sides such as Brentford FC, which has built a model around recruiting players with strong underlying numbers at modest fees and selling them on at a profit [22, 43]. Applied carefully, ML can sharpen strategies of this kind by improving the accuracy of those judgements and lowering the rate of costly errors.

1.3 Problem

A long line of work has tried to compress the data captured in football into a single number that expresses how well a player performed. Such ratings give a convenient analytical handle on the game, yet they can mislead, in large part because of luck. A single fortunate moment can be rewarded out of proportion, shifting how supporters and coaches read a display and even how the player views their own form [21]. The deeper difficulty is that performance is hard to predict at all, since the metrics recorded for every action still leave out much of what shapes the sport. Football is driven by an interaction of physical, mental, tactical, environmental, historical, and cultural factors [35], and with twenty-two players on the pitch one mistake can pull another into a poor decision and turn a promising performance into a disappointing one [64].

What makes this harder is that the factors responsible are rarely present in the datasets used for prediction. Contemporary systems work almost entirely from structured, count-based statistics which describe what a player did but encode nothing of the circumstances that produced it. The

information that would fill that gap, whether a player is returning exhausted, recovering from public criticism, or adapting to an unfamiliar role, tends to live instead in unstructured sources such as news articles, social media posts, and fan forums, where it is difficult to interpret and harder still to fold into a conventional pipeline. Large Language Models (LLMs) offer a way through, since they can read such text and return structured signals that complement the statistical record and reduce the influence of these unobserved factors on prediction [40].

This dissertation therefore studies a feature augmentation framework in which an LLM extracts contextual indicators from pre-match news and encodes them alongside per-match statistics, so that the resulting model captures more of what shapes player performance without flattening the variability that defines the sport.

This framing carries a testable expectation. If the conditions surrounding a fixture can be read from the words written about it, then a model given those conditions should anticipate a player's next performance more closely than one limited to past counts. The hypothesis this dissertation sets out to test can therefore be stated directly:

> Augmenting historical counting statistics with contextual signals that an LLM extracts from editorial media lowers the prediction error for next-match player performance ratings, relative to using those statistics alone.

Testing this hypothesis is organised around three research questions:

1. Can an LLM extract structurally coherent, aligned, and directionally valid contextual signals from unstructured football news articles?
2. To what extent does augmenting per-match statistical features with LLM-extracted contextual signals derived from editorial news articles reduce prediction error for next-match player performance ratings?
3. Does the predictive benefit of article-based contextual features vary across player positions, and which positional roles exhibit the greatest sensitivity to external contextual signals?

1.4 Document Structure

The remainder of this document is organised as follows. Chapter 2 reviews the state of the art across match outcome forecasting, player performance modelling, sentiment and unstructured text analysis, and the use of LLMs, closing with the research gap that motivates the work. Chapter 3 sets out the problem in formal terms, together with the assumptions, the hypothesis, the research questions, and the methodology used to construct the contextual vector and the prediction strategy. Chapter 4 describes the dataset, the feature engineering and contextual vector generation, and the experimental setup, then reports the reliability audit of the extracted signals and the predictive results across positions and model configurations. Chapter 5 draws the findings together, states the limitations and threats to validity, and outlines the work that would carry the framework into new settings.

Chapter 2

State of the Art

This chapter reviews the existing literature across four themes that converge on the research problem defined in Chapter 3: match outcome forecasting, individual player performance prediction, unstructured text and sentiment analysis in sports, and the use of LLMs for structured extraction. The chapter opens with the methodology followed to identify and screen the relevant studies, then addresses each theme in turn, and closes with a consolidated analysis of the gaps that the present work sets out to fill.

2.1 Review Methodology

This section details the Systematic Literature Review (SLR) conducted to identify and evaluate research relevant to the thesis topic. The review followed a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-based framework to ensure methodological transparency and reproducibility.

2.1.1 Strategy

The systematic review process aimed to compile a coherent collection of studies forming the theoretical foundation for this thesis. The process integrated the formulation of search strings, selection of appropriate academic databases, and systematic identification and screening of studies based on predefined criteria.

2.1.2 Knowledge Sources

The review drew upon two major databases: *IEEEXplore* and *Scopus*. These sources were selected due to their extensive coverage of research publications within the field of study.

2.1.3 Search Queries

Search question (Query 1): How have ML and deep learning models been used to predict football team performance or outcomes?

```
(football OR soccer)
AND perform* AND (prediction OR forecasting OR modeling)
AND ("machine learning" OR "deep learning" OR "artificial intelligence")
```

Listing 2.1: Search Query 1

Search question (Query 2): How is athlete or player sentiment, opinion, or emotion analysed in sports research, and how is it linked to prediction or performance analysis?

```
(sport OR football OR soccer)
AND ("sentiment analysis" OR "opinion mining" OR "emotion recognition")
AND (prediction OR analysis)
AND (athlete OR player)
```

Listing 2.2: Search Query 2

Search question (Query 3): How are sentiment, opinion, or emotion signals used to model, analyse, or forecast performance in football and other sports?

```
(football OR soccer OR sport*)
AND perform* AND (prediction OR forecasting OR model* OR analysis)
AND (opinion* OR sentiment* OR emotion*)
```

Listing 2.3: Search Query 3

2.1.4 Retrieval Phase

Each search string was executed in both databases using the title, abstract, and keywords fields. The queries were iteratively adjusted to balance recall and precision, avoiding excessively broad or overly restrictive results. The number of documents retrieved from each query and source is summarised in Table 2.1.

Table 2.1: Query results before exclusion

Database	**Q1**	**Q2**	**Q3**	**Total**
IEEE Xplore	64	43	51	158
Scopus	72	85	479	636
Total	**136**	**128**	**530**	**794**

All retrieved records were imported into *Zotero*, which facilitated duplicate removal and reference management. After eliminating duplicates across databases and queries, a total of **692** unique documents remained.

2.1.5 Screening Phase

To ensure relevance to the scope of this thesis, inclusion and exclusion criteria were defined as follows:

1. **Language**: English.
2. **Publication year**: 2015 or later.
3. **Document type**: Journal article or conference paper.
4. **Availability**: Full text accessible for analysis.
5. **Domain**: Sports context.
6. **Topical relevance**: Addresses at least one of the following: sports performance prediction, unstructured data analysis in sports or machine learning applied to sports analytics.

Applying the first three criteria reduced the number of documents to **613**. Titles and abstracts were then screened to assess relevance to the research focus. This process resulted in **68** potentially relevant studies. After excluding papers without available full texts, **60** studies were retained for detailed analysis in the subsequent sections.

2.2 Match Outcome Forecasting

ML has reshaped sports analytics, pushing the field beyond descriptive summaries toward predictive models that integrate heterogeneous data sources. This shift follows both improved compute and a simple reality: most sports are complex systems, and aggregates alone rarely capture what unfolds on the pitch. Recent literature trends toward hybrid approaches that combine quantitative match and event data with contextual signals, enabling finer estimates of team strength, player impact, and even market dynamics.

2.2.1 Feature Definition

Recent work in match outcome prediction moves away from team descriptors and instead builds higher-dimensional representations that encode context, roster composition, and player contributions. Instead of modelling a match as a static interaction between two teams, modern pipelines reflect that lineups, roles, and tactical intent can change across seasons and within a single campaign.

2.2.1.1 Player-Centric Representations

A recurring issue with team-level modelling is that it compresses heterogeneous player contributions into a small set of averages, obscuring the effect of transfers, injuries, and role changes. To address this, season-specific team vectors are often constructed by aggregating granular player attributes such as passing accuracy, tackle success rate, and goal contribution metrics. These representations can track squad composition more faithfully than historical win-loss summaries alone [1, 28, 39].

This perspective becomes even more relevant in high-scoring sports such as basketball, where outcomes can swing with rotation patterns and the efficiency distribution across lineups. In that setting, individual indicators (e.g., Field Goal Percentage (FGpct), rebounds, and player-level defensive contributions) often provide a sharper snapshot of team strength than long-horizon team averages [45].

Because player-derived team vectors can become large and highly correlated, dimensionality reduction is commonly used to improve stability. Principal Component Analysis (PCA) is frequently applied to reduce collinearity and noise while retaining the dominant variance structure in the aggregated player feature space [1].

2.2.1.2 Spatial and Probabilistic Feature Engineering

While player-centric aggregation improves team descriptions [1, 28, 39, 45], many recent approaches focus on modelling the *value* of actions rather than their frequency. Spatial and probabilistic features capture the idea that the same pass or shot can carry very different meaning depending on where it occurs and what it implies for subsequent states of play. Practically, this changes the feature space from "how much happened" to "how dangerous was what happened, given location and context".

- **Expected Threat (xT):** xT partitions the pitch into zones and assigns each zone a value representing the probability that a possession starting there will eventually lead to a goal. By attributing xT gains to actions, xT separates sterile circulation from genuinely progressive possession, and often provides stronger signals than possession share alone [24].
- **Field Tilt:** Field Tilt measures territorial dominance as a team's share of completed passes in the attacking third relative to the opponent. It emphasises where possession occurs, not just how much possession a team has, and can be more predictive than raw possession metrics [24].

2.2.1.3 Target Engineering and Composite Metrics

A persistent issue in outcome forecasting is the reliance on discrete labels (win, draw, loss) that are noisy, particularly in low-scoring sports such as football where variance and chance play a large role. To reduce label noise and provide a denser learning signal, recent work adopts continuous or composite target variables.

- **Success Score:** The Success Score proposes a continuous target that combines chance quality with execution. Let $S(H)$ denote the score for the home team H; it blends Expected Goals (xG) and Goals $G(H)$, then maps performance to a bounded scale via a sigmoid:

$$S(H) = \frac{1}{1 + e^{-k(M(H) - x_0)}} \quad \text{where} \quad M(H) = G(H) + xG(H) + \beta. \tag{2.1}$$

  The parameters k and β dampen the impact of extreme outliers (e.g., very large scorelines), yielding a smoother target that can improve generalisation compared to training directly on categorical outcomes [37].

2.2.1.4 Tactical and Environmental Context

Feature engineering increasingly encodes tactical intent and external conditions, which can materially affect outcomes in specific sports and competitions.

- **Tactical styles:** Teams can be mapped to tactical profiles such as a high-press, counter-attack, or possession-play team, by grouping event-derived statistics such as passes completed in the opposition half or the average height of the defensive line. These profiles are encoded as feature vectors so models can learn stylistic matchups, not only differences in performance magnitude [37].
- **Environmental factors:** External conditions are often omitted in generic pipelines, even though they can materially influence outcomes across many sports. In cricket, for example, pitch condition categories affect ball behaviour and improve forecasting when included as inputs [49]. The same modelling principle generalises to other domains where the playing surface, weather, altitude, or venue-specific properties systematically shape performance. Binary indicators such as venue familiarity can further capture advantages linked to local conditions and travel effects [6, 49].
- **Temporal form:** To reflect the sequential nature of performance, many approaches compute features over rolling windows. This captures short-term form while limiting information leakage from future matches [20, 37].

2.2.2 Methodologies

Methodological choices have shifted alongside feature design. Classical supervised learning remains strong for smaller datasets and strictly tabular inputs, but much of the recent literature emphasises deep learning architectures to capture non-linear feature interactions, sequential dependencies, and learned representations of entities.

2.2.2.1 Deep Learning Architectures

Neural architectures are increasingly applied to structured sports data to model interactions and time dynamics that static linear baselines struggle to represent.

- **One-Dimensional Convolutional Neural Networks (1D-CNNs):** 1D-CNNs can be adapted to tabular sports inputs by treating each feature vector as a one-dimensional signal. Convolutional filters then learn local correlations between adjacent features, producing feature maps that can outperform manual feature design in practice. Reported results improve outcome prediction relative to logistic regression baselines [62] (see Figure 2.1).

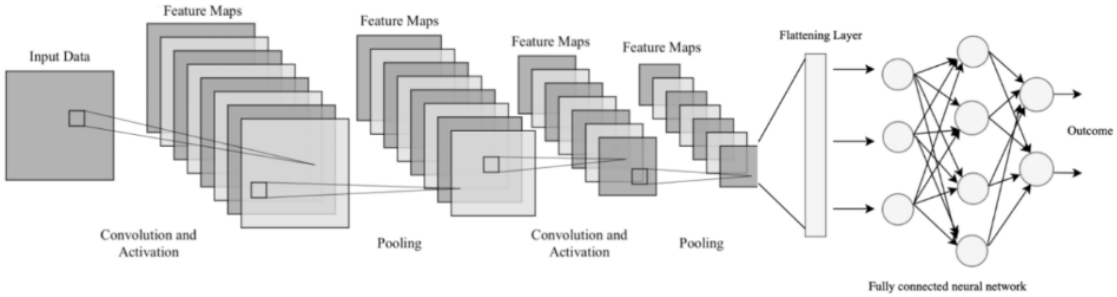

Figure 2.1: Wang and Wang FootballNet convolutional neural network

- **Hybrid (CNN + Transformer):** Hybrid pipelines combine a 1D-CNN, for local feature interaction, with a Transformer module that applies self-attention across time steps. The CNN extracts local patterns from match statistics, then attention weights the relevance of different points in the season, helping represent longer-term changes in team form that purely convolutional or purely recurrent setups may not capture as effectively [20] (see Figures 2.2 and 2.3).

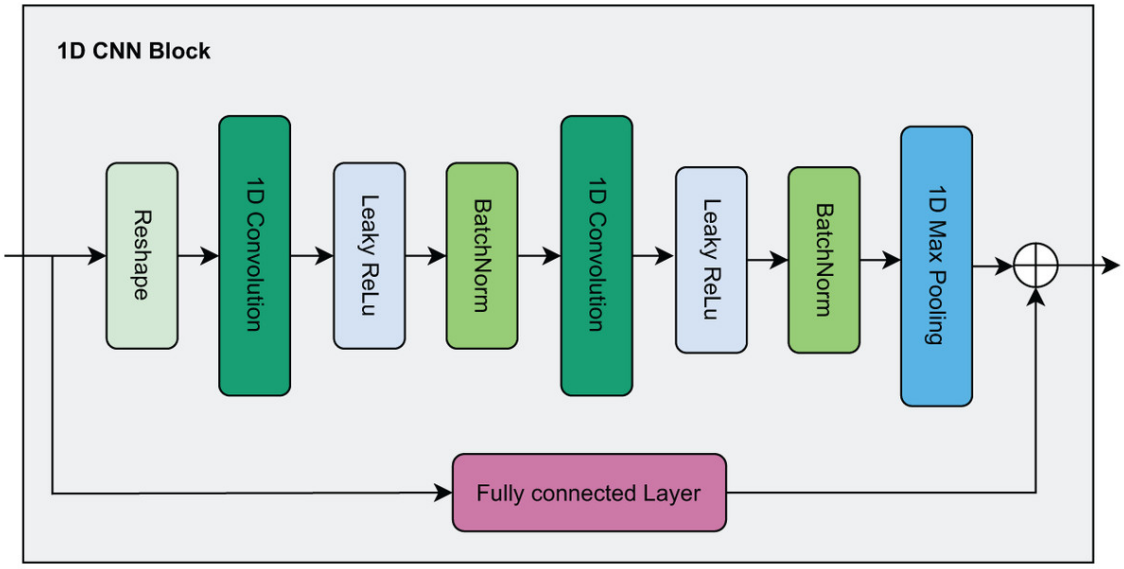

Figure 2.2: Visualisation of the 1D-CNN block in Gao et al.

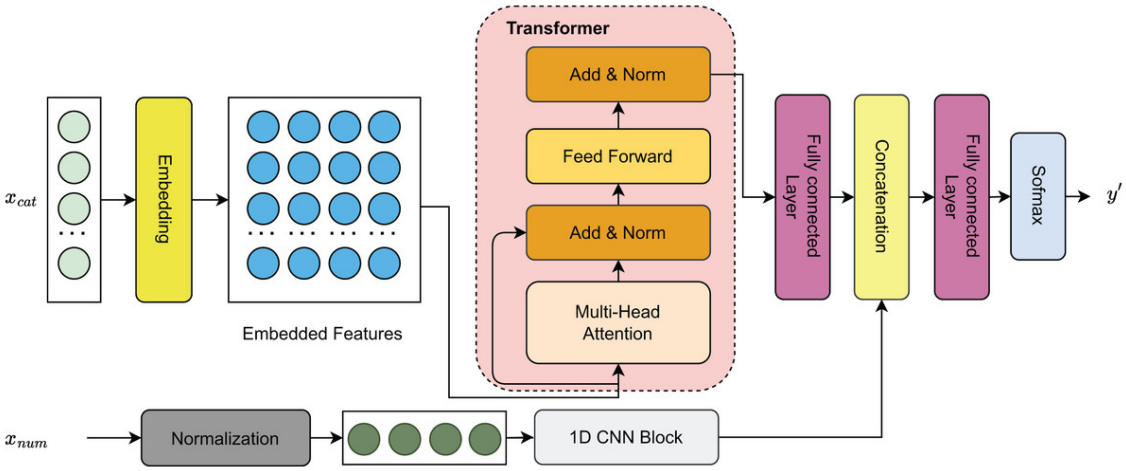

Figure 2.3: Gao et al. model architecture (CNN + Transformer)

- **RNN/LSTM:** Since match outcomes arise from a sequence of performances, recurrent architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memorys (LSTMs) are used to model temporal trajectories. By carrying forward a latent state, these models learn how earlier matches shape later performance, which can outperform approaches that treat fixtures as independent observations [59].

2.2.2.2 Voting Strategies

Even with the rise of deep learning, ensemble methods remain competitive when sample sizes are limited or when inputs remain strictly tabular.

- **Hybrid voting classifiers:** Rather than relying on a single learner, voting ensembles aggregate predictions from complementary models such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) (XGB). Combining models with different error patterns can reduce instability and improve accuracy, particularly in noisy competitions with high player-level variance [6].

2.2.2.3 Advanced Training and Preprocessing Techniques

Beyond architecture selection, recent work emphasises training protocols and data representation choices that improve generalisation and evaluation validity.

- **Entity embeddings:** For high-cardinality categorical variables, embedding layers replace sparse one-hot vectors with dense, low-dimensional representations. This enables models to learn semantic structure from data, where entities with similar behaviour can lie closer in embedding space, improving generalisation to rarer matchups [20].
- **Rolling window validation:** Standard k-fold cross-validation is often unsuitable for temporally ordered match data because it can mix past and future observations. Rolling window

evaluation trains on matches up to time t and tests on the subsequent period $t+1$, then slides forward, preventing look-ahead bias by construction [20].

2.2.3 Outcome Classification Metrics

- **Matthews Correlation Coefficient (MCC):** The MCC is a reliability-focused metric for evaluating binary classifiers, defined as the correlation between the observed classes and the predicted labels. In contrast to accuracy, which may appear high under class imbalance, MCC accounts for all four confusion-matrix terms (True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)) and only yields a strong value when the classifier performs well across both positive and negative cases. The coefficient lies in $[-1,1]$, where 1 indicates perfect agreement, 0 corresponds to chance-level performance, and -1 indicates systematic disagreement. For this reason, MCC is frequently used in sports outcome benchmarks to assess whether a model captures signal beyond majority-class effects [12, 20].

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{2.2}$$

2.2.4 Results and Benchmarks

Empirical results in sports prediction are hard to compare directly because studies differ in:

- **Task formulation**
- **Evaluation protocols**
- **Feature usage**

However, key benchmarks from recent literature are highlighted below.

2.2.4.1 Success Score Framework

Manafzadeh et al. do not train directly on match outcomes, but instead regress a continuous *Success Score* that fuses xG and observed goals as an abstraction for underlying performance quality. They report strong regression fit on a normalised 0 to 10 scale, achieving $R^2 = 0.8592$ [37].

To connect this continuous signal back to decision-style prediction, they discretise the outputs with optimised thresholds ($S > 5.3 \Rightarrow$ win). Using this thresholding approach, they obtain 73.30% accuracy for "win vs. not win" and 75.13% for "lose vs. not lose", supporting the idea that learning a smoother performance target can reduce the instability introduced by noisy categorical labels [37].

2.2.4.2 Deep Learning for Event Outcome Prediction

Gao et al. propose a hybrid CNN–Transformer architecture for three-class prediction (win, draw, loss) and evaluate it under a rolling-window scheme designed to prevent look-ahead bias [20]. Although the reported accuracy is 55.50%, the more informative result is the MCC = 0.275, which indicates predictive value beyond majority-class behaviour. In practical terms, this suggests that the model is not only fitting base rates, but is learning temporal structure that simpler baselines struggle to exploit under the same evaluation constraints [20].

2.2.4.3 Event Data and Real-World Validation

Hassard and Kerr shift the emphasis from aggregate match summaries to event-derived indicators of action quality and territory. Comparing their predictions on real match data from major European leagues (StatsBomb), they show that a linear Support Vector Machine (SVM) trained on these spatially grounded features reaches 65.8% accuracy [24].

This result supports a broader takeaway that appears repeatedly in the literature: where the ball is moved (and what that movement implies) can be more predictive than how often it is moved.

2.2.4.4 Tournament Prediction: The FIFA World Cup Case

Al-Bustami and Ghazal present a distinct evaluation setting because knockout tournaments require a winner, effectively removing draws through extra time and penalties. They therefore frame prediction as a binary classification problem (win vs. loss) and exclude draws from training to mirror the "advance or elimination" structure. Using a RF baseline to assess their player-to-team aggregation strategy, they report an accuracy of 59.38% [1].

Table 2.2: Summary of results and benchmarks reported in the literature

Study	**Label space**	**Reported values**
Manafzadeh et al. [37]	Regression $S \in [0, 10]$	$R^2 = 0.8592$; Acc = 73.30% (W vs. NW); Acc = 75.13% (L vs. NL)
Gao et al. [20]	{Win, Draw, Loss}	Acc = 55.50%; MCC = 0.275
Hassard & Kerr [24]	{Home Win, Draw, Away Win}	Acc = 65.8%
Al-Bustami & Ghazal [1]	{Win, Loss}	Acc = 59.38%
Rodrigues & Pinto [50]	{Home Win, Draw, Away Win}	Acc = 65.26%; Profit margin = 26.78% (RF, EPL)

The approaches reviewed in this section share a structural commitment to team-level outcomes. The target is always a match result or a team-level score, and where individual player data

enters the pipeline it is aggregated into a team vector before prediction. This design suppresses exactly the information that would be needed to forecast how a specific player will perform in a specific match. The player-centric representations of Al-Bustami and Ghazal [1] and the rolling-window form signals [20] show that individual and temporal structure improves team-level forecasts, but neither study follows that logic to its endpoint by predicting individual output directly. A second absence runs across the entire section as no pipeline incorporates unstructured text as an input source. Tactical profiles, environmental factors, and temporal form are all derived from structured event data, leaving the contextual information carried by media reporting, such as injury updates, role changes, and confidence narratives, outside the feature space entirely.

2.3 Player Performance and Impact

Where the previous section targeted team-level match outcomes, the work reviewed here shifts the prediction target to the individual player. The section covers rating systems that define the target variable, models that forecast individual output or development trajectories, and graph-based and dependency-aware approaches that attempt to quantify a player's impact beyond what standard box-score statistics capture.

2.3.1 Player Rating Systems

The prediction target in individual performance forecasting is typically a per-match player rating, a single scalar that compresses the full range of a player's contributions into one number. Platforms publish such ratings after each match, derived from proprietary weighted combinations of event-level statistics.

Several of the approaches reviewed in the following subsections can be read as attempts to build better rating systems. The dual-metric framework of Malikov and Kim, which predicts Expected Goals per game (xGg) and Actual Goals per game (aGg) independently to classify players by scoring efficiency [36], and the Player Impact Metric (PIM), which fuses xG, xT, and ordinal regression weights into a single holistic score [15], both respond to the same limitation. Standard ratings compress context-dependent contributions into a single aggregate that is difficult to decompose after the fact.

2.3.2 Player Performance and Development Forecast

Match outcome forecasting compresses performance into a single team-level result, but individual forecasting targets the specific actions a player is expected to contribute. Earlier work often treated short-horizon production (goals, saves) as the main target, whereas newer work pushes toward forward-looking estimates of *development*, including future quality [60].

2.3.2.1 Feature Definition and Engineering

A central challenge in predicting player output is disentangling a player's intrinsic ability from the opportunities provided by their team and the non-linear nature of human performance trajectories. Contemporary approaches address this by moving beyond raw historical counts to include probabilistic metrics, position-specific contexts, and time-series evolution.

- **Position-aware modelling and dimensionality control:** Because the meaning of a feature changes sharply by role, recent frameworks prefer training separate models per position rather than one global model. Feature filtering is then applied to reduce redundancy and keep the variables that carry the clearest signal for that role, so goalkeeper outcomes are not driven by attacker-centric statistics and vice versa [38].

- **Lagged and historical indicators:** Season-lag features remain the backbone of most pipelines, with prior goal totals repeatedly emerging as strong predictors of future scoring. At the same time, the strength of these lag signals differs across leagues, which aligns with the idea that tactical stability and competitive balance affect how well the past transfers to the near future [39, 60].

- **Dual-metric and efficiency features:** To capture performance quality rather than just quantity, recent work integrates xG alongside actual goals. A dual prediction model has been proposed that forecasts both xGg and aGg independently. By predicting these two metrics separately, a secondary efficiency signal is derived: the divergence between predicted output and expected output allows players to be classified as "Efficient Scorers" (who systematically overperform xG) or "Consistent Performers" (whose output aligns with expectations) [36].

- **Contextual and team-level dependencies:** Individual output is often bounded by team dominance and chance creation, so player models increasingly incorporate team descriptors such as "Team Dangerous Attacks" and conversion-related rates. This hierarchical construction makes the constraint explicit: a player can be excellent and still be capped by low team shot volume or weak chance quality [54, 60].

- **Weighted environmental adjustments:** Comparing players across different competitions introduces bias due to varying league strengths. To mitigate this, constructed features such as `league_weight` and `position_weight` are used. These scalar adjustments allow models to normalise output data from disparate sources, improving the generalisation of regression models trained on multi-league datasets [36].

2.3.2.2 Regression Metrics

Where the prediction target is a continuous value rather than a class label, the metrics used in the outcome forecasting literature no longer apply. Player performance and development models are

instead evaluated with standard regression diagnostics. Mean Absolute Error (MAE) reports the average absolute residual in the same units as the target and provides a directly interpretable measure of typical error. Root Mean Squared Error (RMSE) penalises larger deviations more heavily through its squared term and is preferred when a few large misses are more costly than many small ones. Mean Squared Error (MSE) is the squared form of RMSE and amplifies differences between configurations that are separated mainly by tail behaviour. Several of the studies reviewed below report one or more of these metrics when comparing model families for player-level regression tasks [36, 54, 60].

2.3.2.3 Analysis

Contrary to the assumption that deep learning universally outperforms traditional methods, recent studies establish a "complexity threshold" dictated by player position.

- **Attackers require non-linearity:** For high-variance targets like goal-scoring or transfer value development, ensemble tree methods, specifically XGB, consistently achieve the lowest error rates (RMSE / MAE). This is attributed to their ability to capture non-linear and interaction terms that linear models fail to resolve [54, 60].
- **Defenders and rates favour linearity:** Conversely, strictly linear models have proven superior for goalkeeper metrics (saves/clean sheets) and rate-based statistics (xGg). Research indicates that when defensive features are heavily correlated, Ridge Regression and even standard Multiple Linear Regression (MLR) outperform neural networks. This suggests that defensive output is more largely a function of volume and systematic team structure, that are fundamentally linear, rather than individual brilliance [36, 54].

2.3.3 Player Impact

The forecasting approaches reviewed above predict what a player will produce, but a parallel line of work asks a different question: how much of what happened in a match can be attributed to a specific player. Impact models do not forecast future output, they decompose observed outcomes to quantify each player's contribution, whether through graph-based interaction modelling or through dependency classification schemes that separate individual influence from collective and opponent-driven effects.

2.3.3.1 Graph-Based Representation

Graph-based approaches model match dynamics by representing players as nodes and on-pitch interactions, such as passes, carries, and tackles, as edges. This structural encoding allows player performance to be evaluated through distinct methodological lenses [31, 48]:

- **Graph Neural Networks (GNNs):** By using GNNs, researchers can simultaneously capture the spatial and temporal features of match events to evaluate player performance. This

approach often focuses on predicting changes in xT, a metric that addresses the limitations of traditional statistics by assigning value to facilitators who do not directly impact the scoreline. To achieve this, models integrate a sliding window of the last k events, maintaining the temporal context required to attribute xT changes accurately. As illustrated in Figure 2.4, such architectures typically employ Graph Convolutional Layers (GCLs) to aggregate node information and Multi-Layer Perceptron (MLP) to process edge features, culminating in a Global Mean Pooling layer that generates a unified graph-level embedding [31].

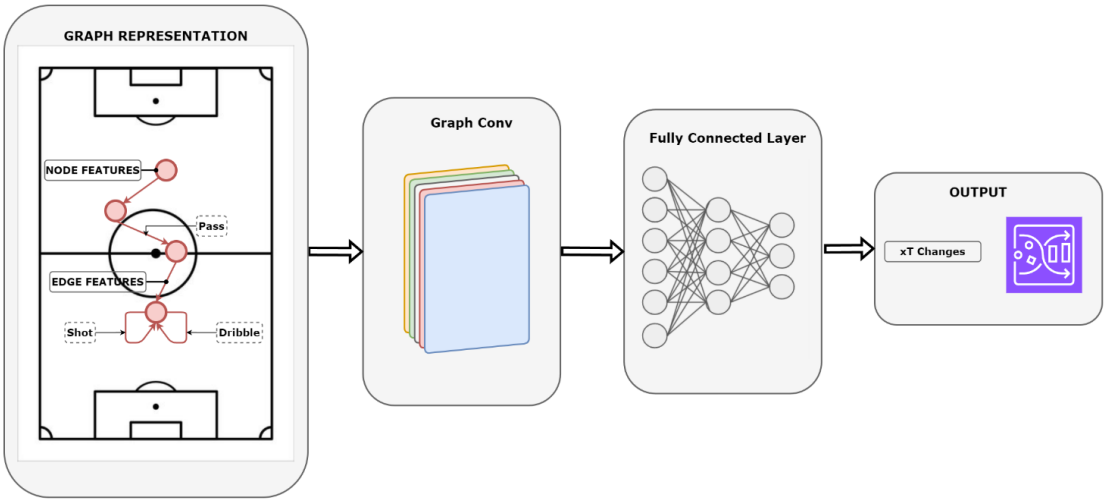

Figure 2.4: Jiang and Cai graph based architecture representation

- **Centrality as a measure of performance:** Principles from classical graph theory have been adapted to quantify player influence. For instance, by applying the Personalized PageRank (PPR) algorithm to a network of player interactions, researchers can compute a centrality score for each player. A higher centrality has been shown to correlate strongly with "Man of the Match" designations, serving as a representative for superior overall performance. Similarly, GNN-based frameworks incorporate metrics like degree centrality to dynamically distribute credit, effectively highlighting "pivotal" players whose contributions are often undervalued by score-centric metrics [31, 48].

2.3.3.2 Player Dependency and Rating

Player Dependency Although not entirely predicting a value for player performance, recent systems attempt to classify match outcomes based on dependency types. Using an Adaptive Back Propagation Neural Network (ABPNN) model, these systems learn non-linear relationships between player statistics and match outcomes to determine whether a result was [33]:

- **Player dependent:** The match outcome resulted from an excellent performance by a specific player

- **Team dependent:** The result was driven by collective team performance, regardless of individual contributions
- **Opponent dependent:** The result was dictated by the opponent's performance

This approach brings a new dimension to performance prediction by correlating team dependency on individual player performance with match victories.

PIM To address the bias towards offensive metrics in traditional ratings, PIM synthesizes match event data, xG, and xT into a single holistic score [15].

- **Construction and Methodology:** The PIM employs ordinal logistic regression to learn the statistical influence of specific actions on match outcomes. Unlike fixed-value systems, this regression assigns dynamic weights w to three primary action categories: Passing, Defense, and Scoring.
- **Contextual Factors:** PIM incorporates contextual elements that traditional metrics overlook:
  1. **Contextual Value:** Utilizes xT grids to value the spatial threat of passes and defensive actions, plus a "Goal Sequence Factor" to credit players involved in the chain of events leading to a goal
  2. **Time Decay:** Applies a temporal factor $\frac{\text{Time}}{\text{Max Time}} + 1$, acknowledging that actions occurring later in a match often carry higher significance
- **Mathematical Formulation:** The component scores are calculated as follows:

$$\text{Pass Score} = (1 + w_{\text{pass}}) \cdot (xT + \text{Goal Seq Factor}) \cdot \left(\frac{\text{Time}}{\text{Max Time}} + 1 \right) \tag{2.3}$$

$$\text{Defense Score} = (1 + w_{\text{defense}}) \cdot (xT + \text{Goal Seq Factor}) \cdot \left(\frac{\text{Time}}{\text{Max Time}} + 1 \right) \tag{2.4}$$

$$\text{Goal Score} = (1 + w_{\text{goal}}) \cdot xG \cdot \left(\frac{\text{Time}}{\text{Max Time}} + 1 \right) \tag{2.5}$$

The overall Player Impact Score aggregates these components:

$$\text{PIM} = \text{Pass Score} + \text{Defense Score} + \text{Goal Score} \tag{2.6}$$

The individual prediction and rating approaches reviewed above operate entirely within the boundary of structured match statistics. Position-aware modelling [38], lagged indicators [39, 60], and team-level contextual features [54] all improve the representation of what happened on the pitch, but none of them encode conditions that are known before the match and that shape how the player is likely to perform, such as accumulated fatigue, recent role changes, or confidence state. The PIM [15] and the dual-metric model [36] refine the target rather than the input, building

better ratings from the same statistical base. What remains missing is a mechanism that augments the pre-match feature space with contextual signals drawn from sources outside the structured record, and that does so in a form the prediction model can consume directly.

2.4 Unstructured Text and Sentiment Analysis

The integration of unstructured data, particularly text, represents a significant frontier in sports analytics. While structured data captures what happened, unstructured data often captures the context, public perception, and qualitative assessment of those events.

2.4.1 Data Sources and Preprocessing Challenges

The efficacy of sentiment analysis is heavily dependent on the data source, and the sports analytics literature mostly splits into two streams: social media and professional editorial writing.

Social media provides high volume, real time reactions, which makes it attractive for tracking fan sentiment during live events. The biggest problem is the noise: slang, abbreviations, jokes, coordinated spam, and bot activity can easily drown the signal, so preprocessing is critical. Work on sports summarisation reflects this reality by using dedicated filtering steps (such as spam detection and user level screening) so the analysis focuses on genuine fan reactions rather than manufactured engagement [2, 4, 52].

Editorial content sits at the other end of the spectrum. It is structured, grammatical, and easier to parse, which often produces cleaner sentiment signals and more reliable contextual cues about team conditions (notably player availability and injuries), even if the volume is smaller than social media [58, 61].

2.4.2 Methodologies

The methods applied to extract sentiment and contextual signals from sports text have evolved along a clear trajectory, from dictionary lookups that score individual words, through recurrent networks that learn sequential dependencies within a sentence, to transformer-based models that attend across the full input and capture compositional meaning. The sections below trace that progression and assess what each generation gains and what it still leaves unresolved.

2.4.2.1 Lexicon and Feature-Based Approaches

Earlier and computationally lighter approaches rely on sentiment lexicons, where tokens are mapped to polarity or emotion labels. In sports settings, resources such as SentiWordNet and the NRC Emotion Lexicon have been used to link words to affective categories, which are then aggregated into match-level or tournament-level indicators [2]. Lexicon-based pipelines remain attractive because the mapping from input to prediction is transparent, but they tend to fail when context inverts meaning, when sarcasm is present, or when domain-specific terms carry sentiment that general dictionaries do not capture [57].

2.4.2.2 Deep Learning, LSTMs, and Affective Computing

Recurrent models, especially LSTMs, address part of the context problem by learning dependencies across a sequence instead of treating a sentence as a bag of words. That matters in sports text, where the sentiment often hinges on a phrase boundary, not on one token in isolation [55].

A concrete application of this principle is the affective computing framework proposed by Liu et al., which combines LSTMs with Word2vec embeddings to extract continuous sentiment signals from post-match reports and feed them directly into a player rating model [61]. The framework consists of two modules. The first is an Affective Computing Module that processes post-match reports through text preprocessing and a Word2vec model to generate word embedding sequences. These sequences are fed into an LSTM network to determine positive or negative confidence, producing an "affective score" ranging from -1 to 1. Simultaneously, the system extracts player names and events using defined lexicons and syntax rules, ultimately producing triples of (Player Name, Event, Affective Score). The second is a Player Performance Evaluation Module that fuses the extracted triples with original match statistics. The original statistic m_k for an event k is modified by summing the affective scores $s_k(i)$ associated with that event:

$$M_k = m_k + \sum_{i=1}^{n} s_k(i) \tag{2.7}$$

Here, M_k represents the modified statistic, which is then input into a linear regression model to output the final player rating [61].

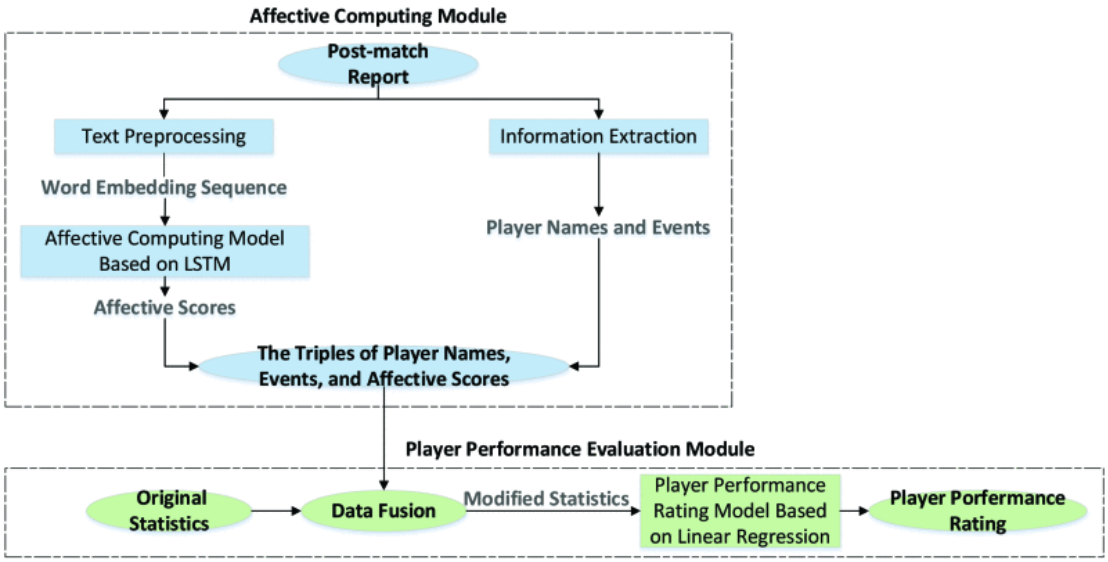

Figure 2.5: Architecture of the improved evaluation method for soccer player performance using affective computing

The system produces a "Player Performance Rating" that integrates both quantitative data (traditional statistics) and qualitative data (affective scores from text), allowing the model to differentiate between events based on their description in professional reports. The method was validated

through two experiments using data from the Chinese Super League (CSL). In the first, the top 20 players identified by the proposed method had a higher total market value (€174.15m) than those identified by a purely statistical baseline (€162m), suggesting the affective adjustment better identifies high-value players. In the second, a Sequential Consistency Score (SCS) metric was used to measure alignment with national team selections; the proposed method achieved lower (better) SCS values across all four positional roles, with the largest improvement for forwards (0.25 against 1.0 for the traditional method) [61].

2.4.2.3 Transformer-Based Models

Recent work increasingly uses transformer-based-models because they handle context through self-attention, which lets the model weigh relationships between words across the entire sequence. These models excel at capturing nuance, long-range dependencies, and the kind of compositional meaning that shows up in both fan discourse and match reporting [4, 58].

The following two studies are discussed as representative examples of current transformer-based-pipelines.

1. **Social Media and the "Wisdom of the Crowd":** This line of work uses the `twitter-roberta-base` family to infer polarity, emotion, and irony from short, informal posts, then turns those outputs into signals that a forecasting model can actually use. In practice, the model predictions are aggregated into engineered features such as Sentiment Ratios (how opinions split across classes), Sentiment Levels (how strong or confident the expressed sentiment is), and Sentiment Imbalance (how far current activity deviates from a historical baseline in both volume and direction). The underlying assumption is simple: crowd talk carries information about momentum, morale, and anticipation that does not show up in official stats, which is why these features have been explored for forecasting tasks in sports settings such as the National Basketball Association (NBA) and Ultimate Fighting Championship (UFC) [4].

2. **Editorial news and hybrid modelling:** Other approaches shift attention from collective opinion to curated reporting, where the language is cleaner and the content often contains concrete context that is hard to extract reliably from social media. Using transformer-based-encoders such as `DeBERTa` and toolkits like `pysentimiento`, these methods extract semantic signals from weekly articles and injury reports, then fuse them with structured match and player statistics in a hybrid predictive pipeline (implemented with 1D-CNN components). Reported results show that the fusion of unstructured data with tabular setup outperforms purely statistical baselines for predicting future player performance in Fantasy Premier League style tasks which directly correlates to in-pitch players performance [58].

Two limitations persist across the text-based approaches reviewed in this section. The first is granularity. Lexicon methods, LSTM pipelines, and transformer-based sentiment classifiers all reduce text to a scalar polarity or a small set of emotion categories, which compresses away

the multi-dimensional contextual information that sports articles carry. A pre-match report may simultaneously convey that a player is fatigued, has been redeployed to a new role, and is playing in a high-stakes fixture, yet a single sentiment score collapses these distinct signals into one number. The affective computing framework [61] moves closer to the individual player level by linking affective scores to specific events, but it operates on post-match reports and therefore cannot serve as a pre-match predictor. The hybrid pipeline [58] does use pre-match editorial content and shows that fusing text with structured statistics outperforms statistical baselines, but it still routes the text through a sentiment encoder rather than extracting structured, multi-dimensional constructs from it. The second limitation is architectural. Every approach in this section uses the language model as a feature extractor whose outputs are consumed by a separate downstream model. None of them asks a language model to reason over both the statistical record and the article text to produce a structured contextual judgement directly, which is the capacity that LLMs have made available and that the present thesis sets out to evaluate.

2.5 Large Language Models

The text-based pipelines reviewed in Section 2.4 treat language models as feature extractors where text enters the model and a fixed-dimensional embedding or sentiment score exits, after which a separate regression or classification model consumes the output. LLMs open a qualitatively different mode of interaction. Rather than mapping text to a vector, a prompted LLM can be asked to read a body of evidence, apply a stated set of rules, and return a structured judgement in a prescribed format.

2.5.1 Reasoning and Extraction via Prompting

When prompted with explicit task definitions and output schemas, LLMs can perform structured information extraction that earlier encoder-based models could not, because the instruction itself specifies what to extract and in what form to return it [63]. In sports analytics this capacity has not yet been applied to the problem of extracting per-player contextual constructs from match reporting. The closest precedent in the reviewed literature is the hybrid pipeline [58], which uses a transformer encoder to produce sentiment features from articles, but the extraction target there is a polarity score rather than a multi-dimensional contextual vector, and the model has no access to the player's statistical record during extraction. An LLM prompted with both the article text and a statistical summary can, in principle, cross-reference one against the other before producing its output.

2.5.2 Failure Modes

Two failure modes are relevant when an LLM is used as a feature generator rather than as a conversational agent. The first is hallucination, where the model produces claims not supported by the input evidence. In an extraction task this manifests as contextual scores that reflect plausible

but fabricated information, for example adjusting a fatigue score based on an injury the article does not mention. The second is inconsistency, where the same input produces different outputs across runs or where semantically equivalent inputs produce different scores. Both failure modes have been documented extensively in the general LLM literature [3, 30] and are expected to appear in any application that treats LLM outputs as quantitative features.

The gap that remains open after this section is whether an LLM can extract contextual signals from unstructured football reporting with sufficient structural coherence, directional validity, and consistency to serve as input features for a downstream prediction model. This is the question posed by the first research question of the present thesis.

2.6 Research Gap Analysis

The preceding sections each closed with an unresolved limitation specific to its theme. Match outcome forecasting operates at the team level and draws on structured data only, leaving individual performance and media-derived context outside the pipeline [20, 24, 50]. Player performance prediction models the individual but restricts its inputs to the structured statistical record, with no mechanism for encoding pre-match contextual conditions. Text-based approaches bring unstructured data into the loop but reduce it to sentiment polarity rather than extracting the multi-dimensional aspects that football reporting actually carries, and they treat the language model as a feature extractor rather than as a reasoning agent.

These per-theme gaps converge on two consolidated shortcomings in the current literature. The first is contextual blindness in the feature space. Advanced metrics such as xT and xG have shifted evaluation from event counts toward action quality, yet the feature sets used for prediction still omit conditions that are known before the match and that shape how a player is likely to perform. Factors such as fatigue accumulation, recent role changes, confidence state, and fixture pressure are discussed routinely in editorial reporting but appear in no structured dataset reviewed here. The affective computing work [61] demonstrated that text-derived signals can improve player ratings, but it operated post-match and reduced text to a single polarity dimension.

The second is the underuse of LLMs as structured extractors. Existing text pipelines apply transformer-based models as encoders that produce embeddings or sentiment labels, which a separate model then consumes. None of the reviewed work asks a language model to read both a player's statistical history and the surrounding media coverage and to return a structured, multi-dimensional contextual vector that the downstream predictor can ingest directly. This is the mode of use that recent LLMs support through prompted reasoning and structured output.

To address these gaps, this work introduces a feature-augmentation framework that combines conventional per-match statistics with contextual signals extracted from pre-match editorial media by an LLM. Chapter 3 formalises this framework and presents an ablation over three contextual configurations that isolates the specific contribution of media-derived evidence.

Chapter 3

Problem Statement

This chapter establishes the formal basis for the proposed research. The problem motivating the study is defined and contextualised against the limitations of existing approaches, followed by the assumptions under which the work operates, the central hypothesis, and the research questions guiding the evaluation. The chapter concludes with an overview of the methodology adopted to address those questions.

3.1 Problem

Predicting individual player performance reduces, in current systems, to a regression over structured, count-based statistics derived from match event data, including metrics such as pass completion rates, shots on target, tackles won, chances created, and xG [24, 45]. These features are retrospective and context-free as they record the observable outcome of a player's actions but encode none of the external circumstances that shaped those actions, which limits how much of the next match they can explain.

The deeper limitation lies in an assumption these pipelines rely on implicitly, that a player deployed in their regular position against a comparable opponent produces statistically equivalent output across fixtures. This assumption fails to account for established contextual modifiers [8]. A player entering a fixture physically exhausted, publicly criticised, or abruptly redeployed to an unfamiliar tactical role carries a different performance prior than the same player entering under stable, high-confidence conditions [7, 13, 51].

The conditions that would correct for this assumption are rarely present in structured datasets and instead reside in unstructured sources such as news articles, social media posts, and fan forums [25, 61]. LLMs provide a mechanism to extract structured patterns from such text and fold them back alongside quantitative data, reducing the influence of these unobserved factors on prediction [29]. Establishing the viability of this route requires first evaluating whether an LLM can reliably extract structurally coherent, aligned, and directionally valid signals from professional editorial text.

A feature augmentation framework is therefore proposed using Liga Portugal data spanning the complete 2024/2025 season and the 2025/2026 season up to fixture 27. Within this framework, an LLM extracts seven domain-specific contextual indicators from pre-match news articles, which are then integrated as structured vectors alongside per-match statistical features.

3.2 Assumptions

The following assumptions are accepted as given throughout this work.

1. **Player ratings reflect individual match performance.** The per-match player ratings provided are treated as a valid and sufficiently accurate proxy for individual on-pitch performance.

2. **Player positions are maintained for the full duration of each match.** Each player is assigned a single positional label per match, and the position-specific models are trained under the assumption that this label reflects the role occupied throughout the entire fixture.

3. **The LLM produces consistent outputs given equivalent explicit evidence.** For a fixed prompt structure and input context, the LLM is assumed to produce contextually equivalent output scores across different player-match cases that present the same explicit evidence pattern. This assumption does not require identical numerical outputs, but rather that the directional and ordinal relationships between scores are preserved.

3.3 Hypothesis

The contextual modifiers described above are absent from the structured features that current systems rely on, yet they are expected to shape how a player performs in the next fixture. If those modifiers can be recovered from editorial media and encoded alongside the statistical record, the resulting model should predict next-match ratings more accurately than one limited to the statistics alone. This research therefore tests the following hypothesis:

> Augmenting historical counting statistics with contextual signals that an LLM extracts from editorial media lowers the prediction error for next-match player performance ratings, relative to using those statistics alone.

3.3.1 Research Questions

To systematically validate the core hypothesis and evaluate the predictive capacity of the proposed framework, the present study addresses the following research questions:

RQ.1 Coherence of Textual Signal Extraction: Can an LLM extract structurally coherent, aligned, and directionally valid contextual signals from unstructured football news articles?

This question establishes the foundation on which the rest of the evaluation depends. Before the extracted signals can be treated as useful inputs, the extraction that produces them has to be shown to be trustworthy, so this question audits three properties of the extracted scores. Structural coherence requires that the outputs are well-formed and complete. Alignment requires that the scores agree with the evidence actually present in the source articles rather than being assigned arbitrarily. Directional validity requires that each score moves toward the correct pole of its scale, so that a condition described in the text raises or lowers the matching feature as intended. If the extraction failed these checks, any predictive result built on it would be uninterpretable, so this question gates the experiments that follow.

RQ.2 Impact on Rating Prediction Error: To what extent does augmenting per-match statistical features with LLM-extracted contextual signals derived from editorial news articles reduce prediction error for next-match player performance ratings?

This is the central question of the work, and it measures whether the contextual signals carry predictive value once they are placed alongside a statistical baseline. The comparison is framed in terms of next-match rating error, so that any reduction can be attributed to the added contextual features rather than to changes in the model or the data. Because these signals can be produced from the statistical history alone, from the editorial text, or from both, this question also separates the contribution of each source, indicating not only whether context helps but where any improvement originates.

RQ.3 Positional Variance and Contextual Sensitivity: Does the predictive benefit of article-based contextual features vary across player positions, and which positional roles exhibit the greatest sensitivity to external contextual signals?

This question refines the previous one by asking whether the effect is uniform across the pitch or concentrated in particular roles. Since a separate model is trained for each position, the predictive gain from contextual features can be examined role by role rather than only in aggregate. The reasoning behind it is that positions whose output depends more on circumstance and narrative are expected to respond more strongly to media-derived context than positions governed by more stable and repeatable patterns. Answering it identifies where the framework delivers the most value and helps explain the mechanism behind any overall effect found in RQ.2.

3.4 Methodology

The methodology adopted to address the research questions is described in the following sections. The first gives a high-level overview of the proposed pipeline and the relationship between its components. The second details the design and generation of the contextual vector, which encodes the qualitative signals extracted from editorial media. The third specifies the prediction strategy, including the position-specific modelling decisions and the architecture chosen to combine sequential and static inputs into a next-match rating forecast.

3.4.1 Overview

The proposed methodology, shown in Figure 3.1, is structured as a pipeline integrating quantitative match data and unstructured editorial media. Structured data is transformed into a standardised statistical baseline, while qualitative indicators representing pre-match player and team states are extracted by an LLM to form contextual vectors. The extraction follows a sequential process: the LLM first receives strictly statistical history to generate a set of baseline contextual scores, then receives pre-match news articles and adjusts those scores against the available textual evidence. This two-step design isolates the media-driven signal and makes the extraction auditable, since the numerical alterations the LLM applies to the statistical baseline can be inspected directly. Varying the evidence supplied at the adjustment step yields the three contextual variants defined in Section 3.4.2.2. Finally, the data is segmented by player position to train separate predictive models that forecast the subsequent match rating.

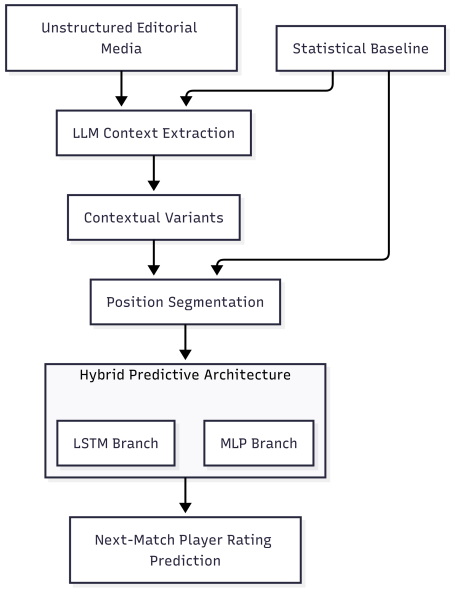

Figure 3.1: Methodology architecture overview

3.4.2 Contextual Vector

The contextual vector is a structured, fixed-dimensional representation of the pre-match state of a player and their team, designed to encode qualitative signals that historical match statistics cannot express. Its construction involves two stages: the definition of the seven feature dimensions and the specification of the three generation variants that isolate the contribution of different information sources. The following subsections address each stage in turn.

3.4.2.1 Conceptual Design

The contextual vector $V_{ctx} \in [0,1]^7$ is a seven-dimensional representation of the pre-match state of a player and their team, where each dimension encodes a qualitative performance signal that

statistical features alone cannot capture. Whereas statistical features record what a player produced in past matches, the contextual vector encodes conditions that are expected to shape their output in the next fixture but leave no direct trace in box-score data.

The choice of a bounded [0,1] interval is deliberate. Rather than allowing unbounded or sign-sensitive scores, a closed unit interval imposes a shared scale across semantically heterogeneous dimensions, which simplifies the model's task of weighting them relative to the continuous statistical features. Each dimension is assigned a fixed directionality: 1 always denotes the pole associated with a specific condition at its maximum intensity, and 0 denotes its opposite. The direction chosen for each feature reflects the football domain interpretation of that condition rather than a learned polarity, ensuring the encoding is interpretable independently of the model. When the available evidence is insufficient to place the player above or below a neutral state, the dimension is set to 0.5, representing the absolute baseline.

Five dimensions describe individual player state. **Fatigue** encodes physical load entering the match, where 1 represents a player who is heavily loaded or exhausted and 0 represents a player who is completely fresh. This dimension is intended to capture the cumulative toll of fixture congestion, travel, or injury return that is invisible in statistics [8]. **Sustainability** encodes output consistency, where a score of 1 reflects stable, recurring form across recent matches and 0 indicates a current average inflated by a single outlier performance within an otherwise weak or erratic run. This metric distinguishes genuinely sustained output from cases where a high rolling average is driven by one exceptional match while surrounding performances remain poor. **Role stability** encodes consistency of deployment, where 1 indicates the player is occupying their usual position and role within the team plan, with habitual minutes, and 0 indicates they have been shifted to a new role or are playing atypically. Players operating outside their primary role or in unusual minutes distributions are known to underperform relative to their statistical baseline [5, 18]. **Confidence** encodes the player's current psychological state and sustained performance level, where 1 represents a player in a positive momentum state, with consistent goal involvement, coach trust, and things consistently going their way across recent matches, and 0 represents a player who is struggling, whether through a poor run, public criticism, or visible loss of morale. **Rate of change** encodes the steepness and recency of form movement, where 1 represents a large, sudden positive shift and 0 represents a sharp decline. This dimension is distinct from confidence. Confidence captures where the player is psychologically and in sustained output terms, whereas rate of change captures how fast and how recently that position shifted.

Two dimensions describe team-level context. **Match stakes** encodes the importance of the fixture to the team at the time it is played, where 1 represents a critical match such as a relegation decider or a fight for a European competition spot and 0 represents a match with no remaining competitive consequence for either side. **Match difficulty** encodes relative opponent strength, where 1 represents the team as a heavy underdog and 0 represents the team as a heavy favourite. While the baseline statistical dataset captures opponent strength through continuous, table-based metrics, **Match difficulty** provides a qualitative assessment of the perceived gap between sides at the moment of the fixture, accounting for media narratives and contextual nuances that raw

standings might miss.

The seven contextual features were selected according to two guiding criteria: each feature had to encode a construct plausibly linked to next-match player rating, and each had to add information not already captured with reasonable fidelity by the statistical features. The first criterion ensures theoretical relevance, the second ensures additive value, since a contextual feature that merely replicates what structured information already expresses would contribute noise rather than new information.

3.4.2.2 Contextual Vector Variants

The generation of contextual vectors is operationalised as a prompt-based extraction task, in which structured match statistics and unstructured news articles are compiled into prompts and processed by an LLM. Three variants were constructed to isolate the contribution of different information sources to the contextual signal. Although each variant produces scores for the same seven features, the information made available to the LLM before scoring differs across variants. This design enables the downstream modelling phase to assess not only whether contextual vectors improve prediction over a statistics-only baseline, but also which information source drives any observed difference in predictive performance.

In the first variant, referred to as Stats-Context, the LLM receives the match context alongside a structured block of a player's match history. The model reasons purely from statistical history to produce the baseline contextual scores, establishing a qualitative layer derived entirely from the same data that the statistical baseline already processes, but encoded through the seven defined constructs rather than as raw numeric features.

In the second variant, Article-Context, the LLM receives the Stats-Context baseline contextual scores as a prior and all news articles mentioning the team within a lookback window before the fixture. The model is instructed to adjust the prior only when the articles provide explicit supporting evidence. This constraint ensures that any deviation from the baseline is strictly driven by the provided media narratives, making this variant the isolated testbed required to assess the coherence and directionality of the LLM's text extraction, thereby directly addressing RQ1 (3.3.1).

In the third variant, Full-Context, the task is the same as in the previous variant but rather than adjusting the values from articles alone, the model cross-references media narratives against the raw performance history to decide whether and how to override the baseline.

Across all three variants, explicit definitions and directionality instructions for each feature are provided to the LLM, rather than requiring the model to infer what each construct should capture. For each feature, the prompt specifies the scale endpoints and their meaning. This constraint was adopted for four reasons. First, without directional anchoring, the LLM may assign internally consistent but semantically inverted scores, producing a signal that is negatively correlated with the intended construct. Second, explicit definitions improve reproducibility because the same evidence maps to the same region of the scale across independent runs when scale endpoints are fixed. Third, consistency across the three variants requires that the same construct is encoded identically in all three prompt configurations. Anchoring directionality guarantees that a `confidence` score

of 0.8 carries the same meaning in Stats-Context as it does in Full-Context. Fourth, the constraint prevents the LLM from assigning its own meaning to each feature, ensuring the resulting scores are interpretable and possible to validate against the intended definitions.

3.4.3 Prediction Strategy

The prediction task itself is formulated as a next-match rating forecast. For each player, the available match history is used to predict the rating of the following fixture, rather than the rating of the current one. This means that the model receives information from match T and is trained to estimate the player's performance in match $T+1$. Considering that the same rating pattern is not shared uniformly across roles, the data was split to train separate position-specific models. Different positions exhibit different statistical profiles and are influenced by different match dynamics, so a single model would force one parameterization across heterogeneous groups. Training separate models by position reduces this mismatch and allows the predictive behaviour to be compared under a more controlled setting (see 2.3.2.1).

For that, a hybrid LSTM+MLP architecture was adopted because the prediction task combines two structurally different forms of input. Match history unfolds as a sequence, so order and recency must be preserved when modelling the temporal component. For this reason, the LSTM was used to encode dynamic match-by-match information, since its gating mechanism is designed to retain relevant context across multiple timesteps and to mitigate the vanishing gradient problem associated with standard recurrent networks [9, 27, 44]. By contrast, fixture-level context is available as a fixed pre-match state and does not have an internal temporal order. A MLP was therefore used for the static branch, as it can model non-linear relations among non-sequential inputs without imposing an artificial ordering [23]. The two branches were fused so that temporal evidence and static context could be learned separately before being combined into a final prediction [47, 56].

This design addresses RQ2 by testing whether contextual information improves predictive performance relative to the statistics-only model, and whether the source of that contextual information changes the gain obtained. It also supports RQ3 by making the same comparisons within each position-specific model, which allows the results to be examined across roles and determines whether the effect of contextual input is consistent or varies by playing profile (3.3.1).

Chapter 4

Evaluation

This chapter evaluates the system developed in the preceding chapters across two dimensions. The first concerns the reliability of the contextual vector extraction pipeline, examining whether the LLM produces directionally valid, evidence-grounded adjustments from editorial media. The second concerns predictive performance, assessing whether the generated contextual features reduce next-match rating error relative to purely statistical baselines.

The following sections describe the empirical basis and design of the evaluation. It begins with a characterisation of the dataset, covering the two Liga Portugal seasons from which all match records were drawn. It then defines the prediction problem and details the preprocessing and feature engineering pipeline applied to the raw data. The contextual vector generation procedure and its three variants are then described, followed by the experimental setup, including the model architecture, data segmentation, and training protocol.

4.1 Data and Feature Engineering

All models evaluated in this chapter draw on a single engineered dataset rather than on the raw match records as collected. This section documents how that dataset was produced, covering the source data, the one-step-ahead forecasting formulation that fixes the target variable, and the preprocessing and feature engineering pipeline that converts raw per-match statistics into the final feature representation used in the modelling stages that follow. The result of this stage is the statistical baseline dataset that the contextual variants defined later in the chapter extend.

4.1.1 Dataset

The dataset covers two Liga Portugal seasons: the complete 2024/2025 season and the 2025/2026 season through fixture 27 of 34. All data was extracted from zerozero [65], which served as the primary data source for this work. Four raw datasets were produced from this extraction, match fixtures, player match statistics, squad compositions, and news articles, whose schemas are documented in Appendix A.1.

The player match statistics table formed the core modelling input, containing 17 134 rows prior to any quality filtering. A single cleaning step was applied at this stage. As the rating constitutes the target variable for all predictive models and rows without it carry no supervisory signal, rows lacking it were removed. Of the 17 134 rows, 455 (2.66%) fell into this category and were dropped, leaving 16 679 records for subsequent processing. The full list of raw statistical columns present in this table is provided in Appendix A.1.

4.1.2 The Prediction Problem

The task addressed in this work is to predict a player's future match performance from information that is available before that match takes place. The rating assigned to each player after a fixture is treated as the target variable, that is, the outcome the model is trained to estimate from the corresponding input features. Because the objective is to anticipate what will happen in the next appearance rather than to describe the current one, the problem is formulated as a one-step-ahead forecasting task, where historical observations are used to estimate the rating of the following match.

This formulation is appropriate because player performance is observed sequentially over time, and each match rating can be interpreted as a realised outcome of an evolving process shaped by prior form, match involvement, and pre-match contextual conditions. Preserving this causal structure requires that only information available before a given match contributes to estimating the rating of that match.

4.1.3 Data Preprocessing and Feature Engineering

The raw statistical records required a sequence of transformations before they could serve as model inputs. The subsections below describe each step in order:

1. Construction and smoothing of efficiency ratios;
2. Normalisation of volume statistics to a per-90-minute basis;
3. Derivation of match context proxies;
4. Encoding of short-term form through rolling aggregations;
5. Truncation of the rating distribution to the [6.0, 8.0] range;
6. Definition of the target variable.

4.1.3.1 Efficiency Metrics and Bayesian Smoothing

Evaluating player execution required the construction of efficiency ratios such as pass accuracy and tackle win rate. The full list of derived efficiency features is provided in Appendix A.2.1. Before these ratios could be computed, a data integrity check identified 48 rows in which the

provider had recorded more successful actions than total attempted actions for the same statistical category. These configurations are logically impossible and would produce efficiency ratios above 1.0 if left uncorrected, silently encoding data entry errors as valid performance signals. All 48 affected rows were dropped, reducing the dataset from 16 679 to 16 631 records.

Standard division then introduced a separate problem: severe variance at small sample sizes. A substitute completing one out of one pass achieved a mathematically perfect 1.0 rate, equating their execution profile with a starter who completed 45 out of 45 passes. To correct for this low-sample distortion, Laplace smoothing was applied as a uniform Bayesian prior to all player efficiency ratios [34]. The transformation adjusted the raw empirical ratio by adding a pseudo-count of 1 to the successful actions and 2 to the total attempted actions:

$$\text{Smoothed Efficiency} = \frac{\text{Successful Actions} + 1}{\text{Attempted Actions} + 2} \tag{4.1}$$

This calculation pulled extreme variations from low-volume samples toward a neutral benchmark of 0.5. As the denominator grew over standard playing cycles, the mathematical effect of the prior diminished naturally, allowing true high-volume indicators to emerge without artificial penalty.

Two columns present in the raw table, `called_up` and `captaincy`, were dropped at this stage as they carry no predictive information relevant to individual match performance. An additional feature, `xg_diff`, was derived as the difference between a player's `xg_total` and their recorded `goals`, encoding the gap between expected and actual scoring output for each fixture.

4.1.3.2 Temporal Normalisation (per-90 Metrics)

Raw counting statistics inherently carry a systematic bias related to playing time. A substitute appearing for 15 minutes cannot theoretically accumulate the same volume of actions as a player who completes a full 90-minute match. If left unadjusted, volume-based features would disproportionately reflect exposure duration rather than actual on-pitch execution, skewing the input representation for the predictive models.

To eliminate this bias, the data preprocessing pipeline normalised all volume-based event metrics into per-90 statistics — the features produced by this transformation are listed in Appendix A.2.2. The calculation scaled each raw statistical count proportionally to the minutes played, projecting the output to a 90-minute equivalent:

$$\text{Metric}_{\text{p90}} = \left(\frac{\text{Raw Count}}{\text{Minutes Played}} \right) \times 90 \tag{4.2}$$

This operation converted absolute volume into a standardised rate. Consequently, all players became directly comparable regardless of their substitution timing or total time on the pitch.

4.1.3.3 Match Context Proxies

Player performance is influenced by the immediate physical demands of the fixture schedule and the broader competitive context of the league standing. To account for these external factors, the feature engineering pipeline integrated proxy variables representing physical load, venue, and relative team strength.

The `days_since_last_match` feature was introduced as a proxy for physical load. This variable measures the calendar gap in days between a player's two most recent consecutive fixtures, encoding the duration available for physical recovery before the next match.

To encode the competitive profile of a fixture, two relative league strength features were computed: `weighted_point_diff` and `weighted_pos_diff`. These represent the difference in accumulated league points and table position between the team and their upcoming opponent:

$$\text{raw\_diff} = \text{value}_{\text{team}} - \text{value}_{\text{opponent}} \tag{4.3}$$

A positive value indicates the team is stronger by that measure and a negative value indicates the opponent is stronger. However, applying these raw differences directly introduces noise in the early stages of the season, when a team's position after three or four matchdays bears little relationship to their actual quality. Consider a side that loses its first three matches, it sits bottom of the table with zero points, and a raw point difference against a mid-table opponent would suggest a large quality gap. In reality, three matchdays of results are too few to distinguish a genuinely weak side from a strong one that has had a difficult opening run of fixtures. To address this, each raw difference was scaled by a `season_progress` fraction defined as:

$$\text{season\_progress} = \frac{T_{\text{current}}}{T_{\text{total}}} \tag{4.4}$$

where T_{current} is the current matchday and T_{total} is the total number of fixtures in the season. The weighted feature is then:

$$\text{Diff}_{\text{weighted}} = \text{Diff}_{\text{raw}} \times \text{season\_progress} \tag{4.5}$$

At matchday 1, `season_progress` equals $\frac{1}{T_{\text{total}}}$, reducing the raw difference to near zero and effectively suppressing the signal until the table becomes informative. As the season advances and standings stabilise, the fraction approaches 1.0, at which point the weighted difference converges to the raw difference and the full competitive context is reflected in the feature.

4.1.3.4 Capturing Short-Term Form and Momentum

Player performance exhibits temporal dependency, commonly referred to as form or momentum. Treating individual fixtures as isolated events discards the influence of immediate past performance. To encode this short-term trajectory, moving aggregations of historical match ratings were integrated into the feature space. Table 4.1 lists the features produced by this step.

Table 4.1: Rolling rating features and their level of aggregation

Feature	**Level**
`rating_roll3`	Player
`rating_roll5`	Player
`rating_roll3_std`	Player
`rating_roll5_std`	Player
`strength_diff_roll3`	Team
`strength_diff_roll5`	Team

The rolling mean features encode the level of a player's recent output, with the two window lengths serving complementary purposes. The 3-match window is more sensitive to immediate shifts in form, while the 5-match window absorbs single-match noise more effectively, together allowing a distinction between a player in a sustained run and one whose recent average is inflated by a single exceptional match.

The standard deviation features encode performance consistency rather than level. Two players with identical rolling means may carry different implications for the next match depending on whether their recent ratings were stable or erratic across that window. These features therefore add information orthogonal to the rolling means, capturing the reliability of recent output independently of its magnitude.

For each fixture, the rolling mean rating of all players in the team is computed over the previous 3 and 5 matches, and the same is computed for the upcoming opponent. `strength_diff_roll3` and `strength_diff_roll5` are then derived as the difference between the team's rolling mean and the opponent's rolling mean. A raw team rolling mean carries no information about the difficulty of the upcoming fixture, the difference between the two sides' recent collective form encodes the competitive context more directly. A positive value indicates the team has been outperforming the opponent in recent fixtures, while a negative value indicates the opponent has been in stronger collective form.

4.1.3.5 Rating Distribution

The raw rating distribution across 16631 valid rows spans the full range from 3.05 to 10.0, with a mean of 6.895, a standard deviation of 0.522, and a positive skew of 0.827 that indicates a longer tail toward higher ratings. As shown in Figure 4.1, the mass is heavily concentrated in the [6.0,8.0] interval, which accounts for 15811 of the 16631 rows (95.0%). The remaining 5.0% is split between 306 rows below 6.0 and 514 rows above 8.0, forming two sparse tails that sit far outside the bulk of observed performances.

These two tails motivate a truncation to the [6.0,8.0] range, with each side excluded for a different reason.

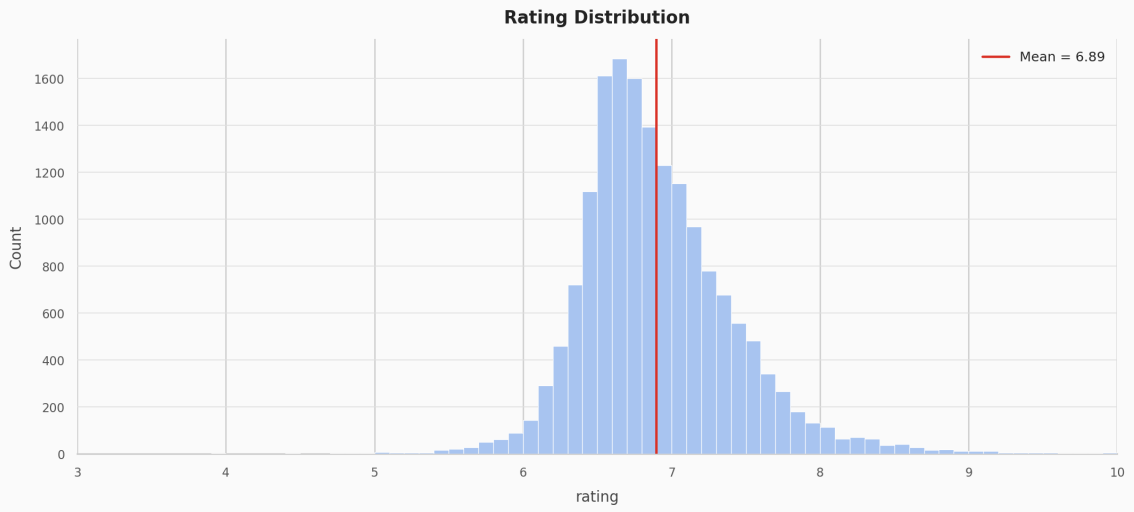

Figure 4.1: Rating distribution

The sub-6.0 tail is not simply rare. It is structurally misaligned with the model's input space. Performances in this range are predominantly the result of disruptive in-match events such as an early red card or a decisive individual error. What these events share is that their occurrence is independent of the pre-match signals available to the model. A player's recent form, fatigue level, or fixture difficulty carries no information about whether such an event will materialise. The consequence is that sub-6.0 ratings cannot, in principle, be predicted from the features the model has access to. Retaining them as training targets would not challenge the model to learn harder patterns but would instead inject irreducible noise into the loss, penalising the model for failing to anticipate events that are unforeseeable from its inputs.

The above-8.0 tail presents a different problem. Ratings in this range do reflect genuine exceptional performances such as a striker scoring twice from open play or a midfielder dominating both phases of the game, and they are not causally disconnected from pre-match features in the same way. The issue is one of sample size. With only 453 rows in [8.0,9.0[and 61 in [9.0,10.0], distributed across four positions and two seasons, there is insufficient volume for the model to learn stable associations between input patterns and these outcomes. Any apparent relationship identified in these bins would be an artefact of the small sample rather than a generalisable signal, and its contribution to the loss would distort the gradient signal for the much larger and more learnable central distribution.

After the cut, 15 811 rows remain. The mean shifts marginally from 6.895 to 6.868, which confirms that the central tendency of the distribution is preserved and the truncation does not distort the target space the model is trained on. More informative are the changes in shape. Skewness falls from 0.827 to 0.460, and kurtosis moves from 3.084 to -0.311. The near-halving of skewness indicates that the asymmetry present in the full distribution was driven by the tails rather than by the structure of the core data, and the sign-flip in kurtosis confirms that the heavy-tail behaviour is removed entirely, with mass now spread more evenly across the retained interval. As shown in Figure 4.2, the resulting distribution covers the full [6.0,8.0] range with sufficient density at every

point, giving the model a well-defined and recoverable prediction target.

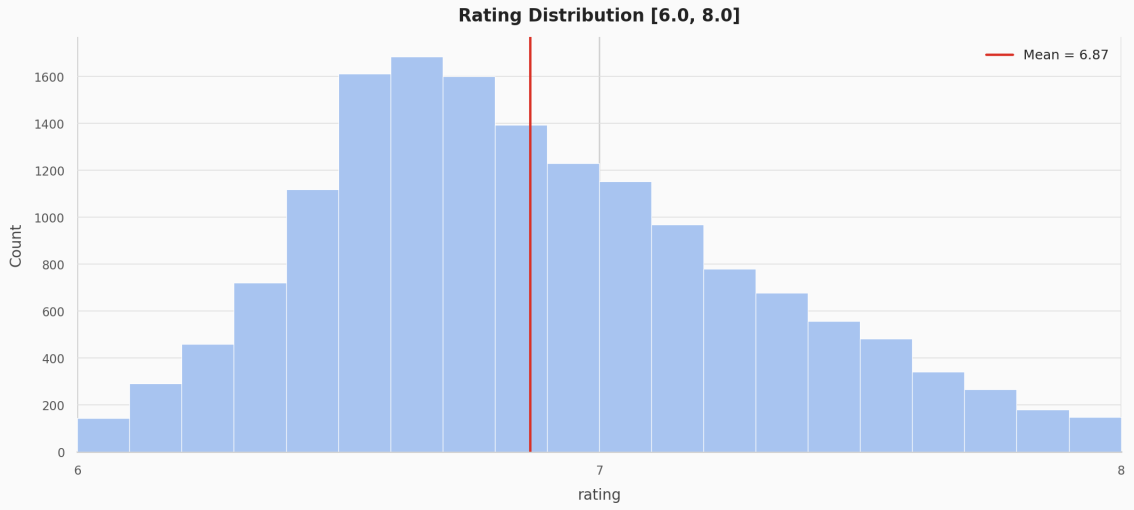

Figure 4.2: Rating distribution after truncation to [6.0,8.0]

4.1.3.6 Target Variable Definition

As established before 4.1.2, the model is trained to estimate the rating of the following fixture rather than the one already observed. To operationalise this, match histories were ordered chronologically within each player-season and the rating at match $T+1$ was shifted backwards to become the label for match T, producing the feature `next_rating`. Season boundaries were enforced before the shift was applied, so that the target feature always refers to an outcome from the same competitive campaign as its input row.

A direct consequence of this shift is that the final match of each player within each season has no subsequent fixture to draw a label from, leaving `next_rating` undefined for those rows. All such rows were dropped, removing 1093 records and yielding a final dataset of 14718 rows. This dataset constitutes the statistical baseline used in all subsequent modelling stages A.2.3.

4.2 Contextual Vector Generation

All contextual vectors were generated using Claude Opus 4.7, accessed through the FCT IAedu API [19]. The design of the seven dimensions and the rationale for the three variants were established in Section 3.4.2, and this section describes how that design was operationalised.

4.2.1 Temporal Boundaries

A central requirement of the extraction pipeline was the prevention of data leakage from the future. When generating the contextual vector anchored at match T, every input passed to the LLM was filtered to exclude any event, statistic, or article dated after the kick-off of T.

Two retrieval windows were defined under this rule and reused across all variants. The statistical history window covered the player's three most recent in-season appearances, $T-3$ to $T-1$. When fewer than three were available the window was truncated, and when none existed the player block was replaced by the fallback statement `No previous matches available this season`. In these cases the LLM was instructed to output 0.5 for all features, representing an uninformed prior at the midpoint of the $[0,1]$ scale. The article window covered the fourteen days preceding kick-off, restricted to articles mentioning the team. This interval was chosen as a design decision to balance recall against the risk of importing outdated narratives.

4.2.2 Article Preprocessing

The article window was phased rather than presented as a flat block. A flat presentation introduced a temporal hallucination risk: an article published ten days before the target fixture might report a player as injured, but the team could have played a subsequent fixture four days later in which that player featured normally. Without explicit temporal segmentation, the model occasionally extracted the older absence as a current one. To mitigate this, the window was bisected at the date of the team's previous match, producing the two blocks described in Table 4.2.

Table 4.2: Phased segmentation of the fourteen-day article window

Block	**Time range**	**Permitted use**
Pre-Previous-Match Context	Up to the team's previous match	Informational only
Current Build-Up	After the team's previous match, up to kick-off of T	Sole source for transient signals

This separation forced the model to reason about the football calendar sequentially rather than treating the window as a single undifferentiated block of text.

4.2.3 Prompt Composition

Each prompt was assembled programmatically from a fixed template. A match context block and a squad list block were common to all variants, while the evidence blocks whose contents varied by variant are described in Section 4.2.4.

The match context block exposed the structured fixture information available before kick-off as a key-value listing, as shown in Listing 4.1. The squad list block enumerated each player in the matchday squad with their internal identifier and position label, formatted as a bulleted list (for example, `[643109] Geovany Quenda (FW)`). It was modelled this way because it defined the set of players for whom scores were to be produced, and it provided an entity resolution anchor against which the model could match free-text player references in the article block.

```
MATCH CONTEXT
- Match ID: 10239952
- Team: Sporting
- Opponent: Nacional
- Date: 2024-08-17
- Venue: away
- season_id: 187713
- matchday: 2
- is_home: False
- opp_id: 27
- team_points: 6.0
- team_league_pos: 1.0
- opp_points: 1.0
- opp_league_pos: 13.0
```

Listing 4.1: Example match context block passed to the LLM for a fixture between Sporting and Nacional

Every prompt closed with the scoring instructions which included the exact feature definitions, directionality conventions, and neutral fallback rule established in Section 3.4.2.1, together with the required output schema.

4.2.4 Variant-Specific Inputs

The three variants share the prompt skeleton and scoring instructions described above, differing only in the evidence blocks used. Table 4.3 summarises which blocks were exposed to each variant and the override rule under which the LLM produced its scores.

In the Stats-Context and Full-Context variants, the statistical history was rendered as a dense comma-separated block per player, as shown in Listing 4.2, where features are collapsed for brevity. The full column list given to the model is provided in Appendix B.1. The article block, structured as the phased window described in Section 4.2.2, was used by the Article-Context and Full-Context variants only. The complete prompt templates from which these per-variant configurations derive are reproduced verbatim in Appendix B.2 and Appendix B.3.

```
Player [639363] Geny Catamo (FW)
date,match_id,minutes,...,strength_diff_roll5
2024-08-09,10239940,76.0,1.0,0.0,7.04,...,-0.00084,-0.00116

Player [500256] Francisco Trincao (FW)
date,match_id,minutes,...,strength_diff_roll5
2024-08-09,10239940,76.0,0.0,0.0,7.96,...,-0.00084,-0.00116
```

Listing 4.2: Example statistical history blocks for two Sporting players, as passed to the LLM

Table 4.3: Evidence blocks and override rules across the three contextual vector variants

Variant	**Stats history**	**Articles**	**Prior scores**	**Override rule**
Stats-Context	Yes	No	N/A	Scores generated from scratch using statistical history only.
Article-Context	No	Yes	Stats-Context	Prior left unchanged unless articles provide explicit supporting evidence.
Full-Context	Yes	Yes	Stats-Context	Prior adjusted by cross-referencing article claims against the statistical record before deviating.

The extraction pipeline processed matches at the team level. For each (team, match) pair in the dataset, a single prompt was submitted covering all players in that team's squad for that fixture. The opposing team was processed independently, meaning two separate inference calls were made per match, one per side, each scoped to the players and context of that team.

To identify which rows were altered by the Article-Context and Full-Context variants, two boolean flags were created, one referencing player-level features and the other referencing team-level features. Figures 4.3 and 4.4 report the share of dataset rows in which at least one contextual score was altered relative to the Stats-Context prior, broken down by the source of that alteration. The most direct measure of overall coverage is the share of rows left entirely unchanged, captured by the Neither bar in each chart. In the Full-Context variant, 4.1% of rows received no modification, meaning 95.9% were touched by at least one score update. In the Article-Context variant this figure rises to 12.9%, so roughly one row in eight carried no article evidence strong enough to override the prior.

Team-level modifications account for the large majority of updated rows in both variants. This is a direct consequence of how two of the contextual features are defined, since `match_stakes` and `match_difficulty` are resolved from fixture-level narratives and therefore take a single value for every player in a given squad for a given match. Whenever evidence shifts either feature, whether through a report on table position pressure or commentary on the strength of the opposition, that update propagates to all players on that side at once, producing many team-level modifications from a single piece of evidence. Player-specific signals, such as individual injury reports or form commentary, are comparatively rare in the article corpus, which is why player-only modifications remain small at 2.5% in Full-Context and 6.6% in Article-Context.

The Full-Context variant fires more broadly than Article-Context across every modification category. The Both share, where team-level and player-level scores were updated simultaneously, is 46.9% in Full-Context against 33.4% in Article-Context, and the Neither share falls from 12.9% in Article-Context to 4.1% in Full-Context. This is expected as Full-Context cross-references article claims against the statistical record before deviating, giving it a richer evidential base from

which to justify updates that articles alone would leave unchanged.

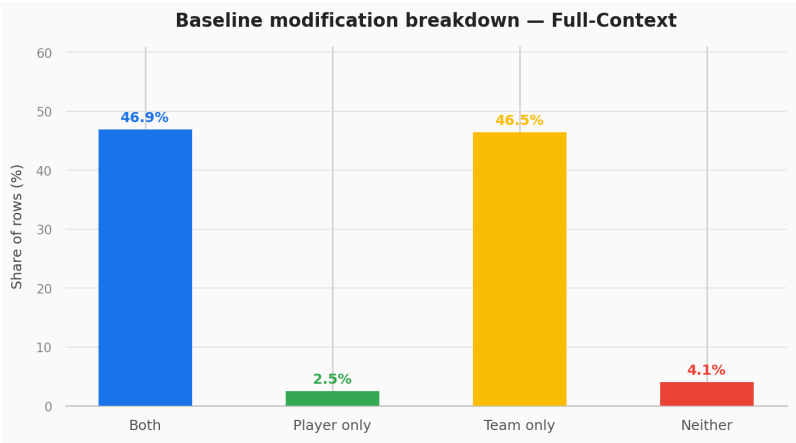

Figure 4.3: Baseline V_{ctx} values modification breakdown for the Full-Context variant

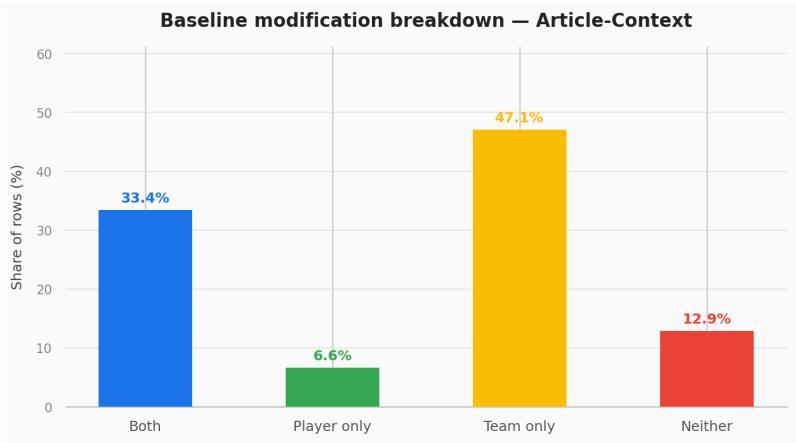

Figure 4.4: Baseline V_{ctx} values modification breakdown for the Article-Context variant

With the modification structure of each variant established, the three enriched datasets were prepared for modelling through two uniform transformations. First, for each V_{ctx} feature at match T, a forward-shifted counterpart ({feature_name}**_next**) was appended by assigning to row T the contextual score from match $T+1$ for the same player and season. This mirrors the construction of `next_rating` and gives the model access to the qualitative pre-match state of the player at the fixture being predicted. Crucially, the contextual vector for $T+1$ is itself generated from information available strictly before $T+1$ kicks off, as established in Section 4.2.1, so no temporal leakage is introduced as the model receives a description of conditions entering $T+1$, not an outcome of it.

4.2.5 Reliability Audit

The contextual vectors are generated entirely from LLM inference over unstructured text and statistical history, with no ground-truth labels against which to validate them automatically. To

assess whether the article-driven adjustments are directionally correct, grounded in genuine textual evidence, and appropriately sized relative to that evidence, a manual audit was conducted. For a set of selected player-match cases, the same articles available to the LLM were read by a human reviewer and each adjustment was evaluated against the source material.

The audit was restricted to the variant in which articles are the sole adjustment mechanism, with no confounding from additional statistical inputs. The other article-conditioned variant draws on both articles and match history jointly, making it harder to attribute a given adjustment to article content specifically. Restricting the audit to the purer configuration provides the cleanest available test of whether the LLM can extract valid, well-calibrated signals from unstructured text alone.

Because team-level features are identical for all players on the same side within a given fixture, reviewing them at the player-match level would be redundant; only the player-level contextual features were therefore used. Player-level features encode more granular, player-specific information that is less anchored in structured statistics, making them the more appropriate target for manual validation.

The 30 player-match cases with the largest absolute deviation ($|\Delta|$) across the five player-level V_{ctx} features were selected for review, where:

$$|\Delta| = |\text{article value} - \text{baseline value}| \tag{4.6}$$

These cases represent the strongest interventions made by the model and are therefore the most informative for assessing whether large overrides are grounded in genuine textual evidence. The complete list of selected cases, alongside their baseline and adjusted feature values, is provided in Appendix C.

Each adjustment was assessed along three criteria, summarised in Table 4.4. **Direction** captures whether the sign of the change was appropriate given the article content, for example, lowering `role_stability` after a confirmed absence is directionally correct, while lowering `fatigue` for a player who is injured is not. **Evidence support** captures how directly the articles justify the specific feature that was changed, distinguishing between explicit one-to-one mappings and indirect or absent support. **Magnitude** captures whether the step size was proportionate to the strength of that evidence.

Table 4.4: Scoring scheme for the three audit criteria. Direction is given double weight in the composite score.

Criterion	**Levels**	**Score**	**Weight**
Direction	Correct / Incorrect	1 / 0	$\times 2$
Evidence support	Strong / Weak / Absent	1 / 0.5 / 0	$\times 1$
Magnitude	Reasonable / Over or Under	1 / 0.5	$\times 1$

The three weighted scores are combined into a composite row score in $[0, 1]$:

$$\text{score} = \frac{2 \cdot \text{direction} + \text{evidence} + \text{magnitude}}{4} \tag{4.7}$$

Direction is given double weight because a directionally incorrect adjustment misrepresents the feature regardless of how evidence and magnitude are otherwise assessed. Cases in which the articles contained stale information, such as an outdated injury that had since resolved, were flagged separately and treated as upstream data quality issues rather than model reasoning errors.

4.3 Experimental Setup

The description and rationale behind the chosen architecture can be found in Section 3.4.3. The implementation details are depicted below.

4.3.1 Position-Specific Modelling Scope

Each outfield position is modelled independently, a practice with precedent in football performance prediction 2.3.2.1. The statistical features that characterise a defender's contribution are largely uninformative for forwards, and pooling the two into a single model would force the same parameter space to explain performance patterns governed by different underlying processes.

The dataset contains 14718 rows after the $[6.0, 8.0]$ rating cut, distributed across positions as shown in Table 4.5. Goalkeepers account for only 907 rows, too few to support reliable generalisation across training, validation, and test splits, and are therefore excluded from the modelling scope. A model trained on a sample of this size would reflect the particularities of the specific matches available rather than any generalisable pattern in goalkeeper behaviour. The three outfield positions each provide between 4226 and 4933 rows, sufficient to support independent modelling.

Table 4.5: Row counts and dataset share by position

Position	**Rows**	**Share (%)**
GK	907	6.2
DF	4652	31.6
MF	4933	33.5
FW	4226	28.7
Total	14718	**100.0**

The rating distributions across the three modelled positions are broadly similar, with means ranging from 6.83 (Forwards (FWs)) to 6.91 (Defenders (DFs)) and standard deviations between 0.39 and 0.42, as shown in Figure 4.5. This consistency confirms that position-specific modelling does not fragment the target variable but rather allows the feature space to be adapted independently for each role.

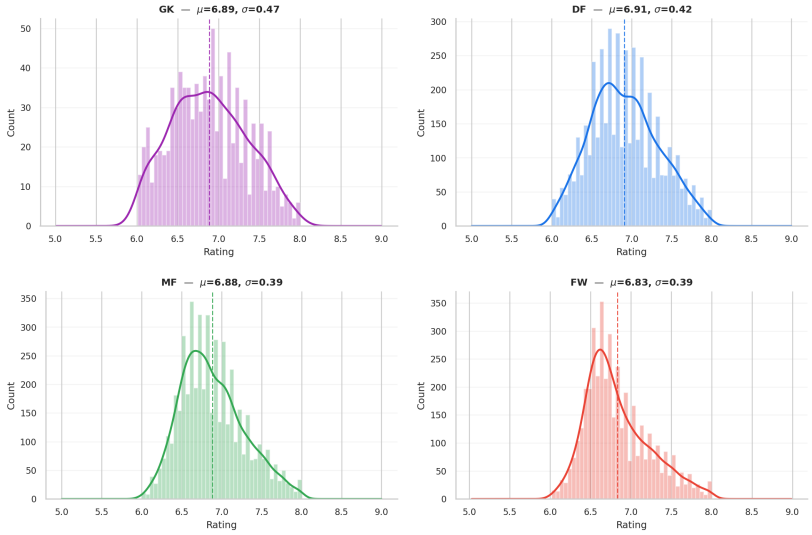

Figure 4.5: Rating distribution per position

4.3.2 Data Segmentation and Feature Groups

Before partitioning the features into architectural branches, position-specific pruning was applied to eliminate structural zeros and uninformative variables. Features containing more than 90% zero values were flagged as candidates for removal. Within this subset, any feature exhibiting an absolute Pearson correlation with the contemporaneous match `rating` below 0.05 was dropped. The correlation was computed against `rating` rather than `next_rating`. The threshold was set deliberately low to evaluate whether a feature carried any empirical association with match performance for that position, rather than testing its forecasting capacity for the following fixture.

The columns removed per position and the resulting schema dimensions are presented in Table 4.6.

Table 4.6: Feature pruning results per position

Position	Rows	Columns dropped	Final columns
FW	4226	8	70
MF	4933	8	70
DF	4652	10	68
GK	907	23	55

The removed columns primarily consisted of Goalkeeper (GK)-specific statistics such as `gk_saves_total_p90` and `gk_high_claims_p90` that were dropped uniformly across the FW, Midfielder (MF), and DF datasets because they are structurally zero for outfield players. Although a small number of role-specific columns were also removed from these outfield datasets, the largest reduction occurred in the goalkeeper subset where 23 columns were discarded because the baseline schema was heavily oriented toward outfield actions.

Following pruning, the retained feature set was partitioned into two non-overlapping groups that map directly onto the two branches of the model architecture described in Section 3.4.3.

The dynamic feature group was routed to the LSTM branch and comprised all per-match statistics that vary across appearances. These fall into three conceptual families: Laplace-smoothed efficiency ratios that capture how often a player's attempts succeed, per-90 rates that express volume of action independently of minutes played, and the residual raw event counts retained after position-specific pruning. Together, these variables describe what a player produced in a given match and therefore carry the per-timestep signal the LSTM was designed to model.

The static feature group was routed to the MLP branch and described the upcoming fixture rather than past performance. It combined fixture-level descriptors, namely venue, the number of days since the last match, matchday index, and the weighted league point and position differences between the two teams, with, depending on the variant, the contextual vector V_{ctx} scores for the current match and their forward-shifted next-match counterparts. These variables are constant within a single prediction instance, since they characterise the fixture at match $T+1$ rather than any event occurring during the historical window $[T-4, T]$.

The rolling rating features were excluded from both branches. These variables were used upstream to inform the LLM during contextual vector generation but were withheld from the predictive model itself, since passing them to the LSTM would have introduced redundancy with the raw rating sequence already present in the dynamic input. Under that configuration, the temporal branch would have observed both the individual match ratings and a pre-aggregated summary of those same ratings, giving it a shortcut to the trajectory dynamics it is meant to learn directly from the sequence.

After position-specific pruning, the size of each feature group differs slightly across outfield positions, as summarised in Table 4.7. The complete list of variables assigned to each branch, for every position, is reported in Appendix A.3.

Table 4.7: Number of input features per position after pruning, separated into dynamic and static groups. Contextual vector (V_{ctx}) features are reported separately because they are only active in the three contextual variants.

Position	**Dynamic features**	**Static features**	V_{ctx} **features**
FW	50	5	14
MF	50	5	14
DF	48	5	14

All numerical input features were standardised to zero mean and unit variance using `StandardScaler`. The scaler was fitted exclusively on the training partition of each position–variant combination and then applied without refitting to the corresponding validation and test partitions, so that no information from later matches could leak into the training distribution through the normalisation statistics. The target variable `next_rating` was kept in its native $[6.0, 8.0]$ scale.

4.3.3 Temporal Train/Validation/Test Split

Because the prediction task is forward-looking, the dataset was partitioned chronologically rather than at random. A random split would intermix past and future observations within the same fold, allowing the model to implicitly learn from matches that post-date the fixture being predicted.

Within each position-specific dataset, rows were sorted by match date. The chronologically latest 20% of rows were held out as the test set. From the remaining 80%, the latest 15% were assigned to validation and the earlier rows to training. The resulting split sizes are reported in Table 4.8.

Table 4.8: Chronological train/validation/test split sizes by position

Position	Train	Validation	Test	Total
FW	2874	507	845	4226
MF	3355	592	986	4933
DF	3164	558	930	4652

By construction, the test set contains only matches played after every training and validation match. Reported test performance therefore reflects out-of-sample generalisation to later fixtures rather than interpolation across a mixed timeline.

4.3.4 Network Parameterization and Training Dynamics

To ensure consistent evaluation across different feature sets, the network architecture and training hyperparameters were standardised across all modelled outfield positions and variants. The configuration details are summarised in Table 4.9.

Table 4.9: Summary of network architecture and training hyperparameters for outfield models

Parameter	Configuration
Temporal Branch	2-layer LSTM (128 hidden units)
Static Branch	2-layer MLP (128, 64 hidden units)
Regularization	Dropout (0.25), Layer Normalization
Activation	ReLU
Optimizer	AdamW
Learning Rate	10^{-3}
Weight Decay	10^{-4}
Batch Size	256
LR Scheduler	ReduceLROnPlateau (factor 0.4, patience 13, min 10^{-6})
Gradient Clipping	1.0 (max norm)
Loss Function	Root Mean Squared Error (RMSE)

The hybrid architecture processes the dynamic and static feature groups through distinct branches before fusion. The temporal branch employs a two-layer LSTM with a hidden state of 128

units, applying a 0.25 dropout rate between layers. To stabilise the variance before fusion, the final hidden state of the sequence is subsequently passed through a Layer Normalization module. Concurrently, the static branch processes the fixture-level and contextual features through an MLP comprising two hidden layers of 128 and 64 units. Each linear transformation in this branch is followed by Layer Normalization, a ReLU activation, and a 0.25 dropout mask.

Following independent processing, the 128-dimensional temporal representation and the 64-dimensional static representation are concatenated. This fused vector is passed through a final prediction head, consisting of a linear layer that halves the dimensionality, a ReLU activation, dropout, and a terminal linear projection, to yield the scalar `next_rating` estimate.

These dimensions and hyperparameters were not derived from an exhaustive search. Instead, they were established early as reasonable defaults for a dataset of this size. Given that the per-position training sets contained only a few thousand sequences, deeper or wider networks would likely overfit before yielding measurable gains. Furthermore, holding the architecture and training dynamics constant across positions and variants ensured that any observed differences in test performance could be attributed directly to the input feature sets rather than to variant-specific tuning.

Optimisation was carried out using AdamW, with the learning rate dynamically adjusted by a `ReduceLROnPlateau` scheduler monitoring the validation loss. To prevent the exploding-gradient behaviour that can arise when backpropagating through time over sequences of unequal length, gradients were clipped to a maximum norm of 1.0.

To mitigate the risk of overfitting and to enforce a uniform termination policy across all neural network variants, an early stopping protocol was integrated into the training procedure. Following the completion of each epoch, the empirical RMSE was calculated over the validation set. If the validation loss failed to decrease by a minimum absolute threshold of 5×10^{-4} relative to the lowest loss recorded up to that point, a stagnation counter was incremented. The training loop was terminated automatically if 40 consecutive epochs elapsed without satisfying this improvement threshold. Upon early termination, the network parameters were automatically reverted to the weights associated with the optimal validation checkpoint, ensuring that final test-set evaluation was executed near the optimal generalisation state of each model.

The training objective was a differentiable RMSE loss between the predicted and observed `next_rating`. This metric was selected because it penalizes large deviations more heavily than smaller ones. This property is particularly relevant in the compressed [6.0, 8.0] target range, where a one-point error corresponds to a qualitatively different performance assessment.

The network exhibited a tendency to concentrate its capacity on the dense central region of the rating distribution and ignore the informative deviations in the tails. To counteract this, a dynamic sample weighting scheme was integrated into the loss function. Training targets were grouped into bins of width 0.1. For each sample i falling into bin k, an inverse-frequency weight was computed as $w_i = N/C_k$, where N is the total number of training samples and C_k is the number of samples in bin k. To preserve the overall magnitude of the loss, these weights were normalised by their mean, yielding the final per-sample weight:

$$\hat{w}_i = \frac{w_i}{\frac{1}{N}\sum_{j=1}^{N} w_j} \tag{4.8}$$

These normalised weights were applied inside the RMSE computation. Consequently, errors on rare, exceptional performances contributed proportionally more to the gradient than equally sized errors on typical matches.

Even with this weighting scheme applied, early experiments showed that models rapidly converged toward predicting the global mean rating for every player, effectively behaving as a constant baseline during the initial epochs. To prevent wasting optimisation steps on this trivial solution, the bias term of the final linear layer was initialised to the mean of the training targets. Under this initialisation, the model started at the training-set average by construction, allowing it to immediately focus on learning residuals, specifically the per-player, per-match deviations from that mean induced by the input features.

Regarding the temporal sequence construction, the dynamic input to the LSTM consisted of up to five previous matches for the same player, ending at the anchor match T. Shorter histories arose naturally for players entering early in a season or following a season boundary reset. Because sequences within a batch possessed variable lengths, shorter histories were zero-padded to the maximum length of five timesteps. Feeding these padded tensors directly into the LSTM would allow the zero rows to influence the recurrent state, thereby biasing the final hidden representation toward an artificial baseline. To avoid this, sequences were processed using PyTorch's `pack_padded_sequence` before the recurrent forward pass. This utility masked the padded timesteps, ensuring that the LSTM only updated its hidden state over genuine match observations and that the extracted hidden state reflected solely the real portion of each player's recent history.

4.3.5 Baselines and Model Variants

To isolate the predictive impact of the generated contextual features, five distinct model configurations were evaluated across each outfield position. Because the generation mechanisms for the contextual vectors were established previously in Section 3.4.2.2, this section defines how those vectors were combined with the standard quantitative features to form the final predictive models.

The configurations were structured as an ablation study. The underlying network architecture and training dynamics remained identical across all neural variants, with variation restricted exclusively to the input feature space. The evaluated variants are summarised in Table 4.10.

The first configuration, designated as the Average Baseline, served as a naive benchmark. It predicted the target rating by calculating the arithmetic mean of the player's previous five match ratings, establishing a floor to verify whether the neural models learned temporal patterns beyond a simple historical average.

The Statistical Baseline used the hybrid architecture but restricted the input strictly to the engineered quantitative features defined in Section 4.3.2. Under this configuration, the static branch received only the fixture-level context, completely isolating the network from any qualitative representation.

Table 4.10: Summary of model configurations and their respective input feature spaces

Model Variant	**Feature Composition**
Average Baseline	Non-neural heuristic predicting the arithmetic mean of the player's recent match ratings.
Statistical Baseline	Dynamic match statistics and static fixture context. Excludes all V_{ctx} features.
Stats-Context	Baseline features appended with V_{ctx} derived purely from statistical history.
Article-Context	Baseline features appended with V_{ctx} adjusted strictly by news article evidence.
Full-Context	Baseline features appended with V_{ctx} generated by cross-referencing articles against statistics.

The remaining three models integrated the contextual vectors. The Stats-Context, Article-Context, and Full-Context models extended the Statistical Baseline by appending their respective contextual vectors to the static feature group. Comparing the prediction errors across these configurations provides the framework to answer the core research questions. Specifically, evaluating these variants in sequence isolates the specific contribution of unstructured text, distinguishing the predictive value of imposing a qualitative structure on existing statistics from the value of the media narratives themselves.

4.3.6 Evaluation Metrics

Predictive accuracy was measured with three continuous regression metrics computed over a common test split, so that every configuration was scored on the same match instances. A permutation-based significance test then assessed whether the resulting differences between configurations reflected a real change in predictive behaviour or fell within the range expected from chance.

4.3.6.1 Predictive Accuracy Metrics

To quantify the predictive accuracy of the position-specific models, three standard continuous regression metrics were calculated: MSE, RMSE, and MAE. The primary optimisation objective and evaluation criteria relied on RMSE, which was selected because it square-penalizes larger prediction errors more heavily than smaller deviations. This mathematical property is relevant given the compressed $[6.0, 8.0]$ target range of the `next_rating` variable, where small absolute differences correspond to distinct performance categories. Conversely, MSE and MAE were recorded as secondary diagnostic indicators to isolate the variance of the errors and provide an easily interpretable measure of average absolute residual magnitude. These metrics are mathematically defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{4.9}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{4.10}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{4.11}$$

where N represents the total number of match instances within the evaluated data split, y_i denotes the true observed performance rating for instance i, and $\hat{y}_i$ represents the corresponding rating predicted by the model.

4.3.6.2 Statistical Significance

To assess whether an observed RMSE difference between two model configurations can be attributed to chance, a paired Monte Carlo permutation test was applied to every pairwise comparison between variants and baselines. Because all configurations predict the same test rows, each test match yields one pair of squared errors, one per model. Under the null hypothesis that the two configurations are interchangeable, the assignment of each pair to the two models is arbitrary. The test therefore swaps the two errors of each match with probability 0.5, recomputes the RMSE difference over the relabelled errors, and repeats this procedure 50 000 times. The p-value is the fraction of permutations producing an absolute difference at least as large as the observed one, and a difference is treated as statistically significant at $p < 0.05$. Comparisons in which no permutation reached the observed difference are reported as $p < 0.0001$.

4.4 Contextual Signal Reliability

Following the audit design established in Section 4.2.5, a manual audit was conducted on the thirty player-match cases with the largest absolute score deviations from the Stats-Context prior, assessing whether the article-driven adjustments are directionally correct and grounded in genuine textual evidence.

4.4.1 Audit Results

A directional accuracy of 80.0% was achieved, with a mean per-case score of 0.84 out of 1.0. This audited sample was positionally skewed, containing 20 rows corresponding to forwards (66.7%), 7 to midfielders (23.3%), and 3 to defenders (10.0%).

Directional accuracy varied by feature, with `confidence`, `role_stability`, and `rate_of_change` all reaching 100% within the reviewed sample. This perfect accuracy indicates that the correct sign of change was consistently identified for form trajectory, squad role continuity, and short-term momentum. Conversely, `fatigue` recorded a lower directional accuracy of

50%, successfully identifying only 6 correct cases out of 12. No modifications were recorded for `sustainability` in the reviewed sample, which aligns with the nature of the feature, as explicit textual evidence for abstract constructs is rarely found within news articles.

Evidence quality was rated as strong in 76.7% of the reviewed rows, indicating that the articles contained explicit support for the modified feature. Although the remaining 23.3% of rows carried weak support, zero cases were rated as entirely absent of evidence. This complete lack of unsupported cases confirms that unsupported interventions were not generated within this deviation sample.

Outdated or insufficiently refreshed information affected 56.7% of the reviewed rows, yet this exposure was heavily concentrated in specific features. For instance, incorrect content associated with stale news was flagged in 11 of the 12 `fatigue` cases, amounting to an error rate of 91.7%. In contrast, `confidence` recorded zero cases flagged out of eight. Between these extremes, intermediate levels of incorrect information were observed for `rate_of_change` at 66.7% and `role_stability` at 57.1%.

Magnitude calibration exhibited a conservative bias, demonstrated by under-adjusted changes occurring in 16.7% of rows while over-adjusted changes occurred in only 6.7%. Furthermore, an intervention delta of 0.4 to 0.6 emerged as the most common adjustment range, covering 18 of the 30 reviewed rows. Within this specific 0.4 to 0.6 interval, a directional accuracy of 88.9% was achieved.

To contextualise the positional skew observed within the audited sample against the broader modification structure of the Article-Context variant, the raw modification counts across the full dataset are reported in Table 4.11, broken down by position and feature direction.

An asymmetric modification pattern is visible across all four positions, as upward adjustments outnumber downward adjustments for every feature. The largest total volume of modifications is absorbed by `confidence`, with upward revisions dominating this feature at every position. In contrast, near-zero modification counts are recorded for `sustainability` across all positions, a finding that is consistent with its absence from the high-deviation audit sample.

4.4.2 Audit Analysis

The 80.0% directional accuracy and 0.84 mean row score demonstrate that directionally valid signals are successfully extracted from unstructured text in the majority of evaluated cases. Furthermore, the 100% accuracy achieved on `confidence`, `role_stability` and `rate_of_change` aligns with the definition of these features, given that both map onto narrative constructs explicitly and repeatedly addressed in football journalism. Because factors such as coach trust, goal involvement, and positional changes are reported in explicit terms, this dense coverage provides direct textual evidence for the model to anchor its adjustments.

This interpretation is reinforced by the asymmetric modification counts visible in Table 4.11. The predominance of upward `confidence` adjustments across all positions reflects the framing of pre-match build-up articles, in which squads in competitive contention receive coverage that foregrounds positive form narratives. Consequently, the evidence arriving via the article window is

Table 4.11: Per-position feature modifications in the Article-Context variant. Each cell reports the number of dataset rows in which the feature was increased (↑) or decreased (↓) relative to the Stats-Context prior. The Total Modifications row gives the column sum of all feature-level changes, and therefore counts modification events rather than distinct rows.

Feature	**Dir.**	**GK**	**DF**	**MF**	**FW**	**Total**
Total Modifications		633	2459	3200	3261	9553
`fatigue`	↑	39	360	448	311	1158
	↓	13	134	139	75	361
`sustainability`	↑	0	6	33	51	90
	↓	2	1	1	1	5
`role_stability`	↑	66	489	724	601	1880
	↓	8	158	139	186	491
`confidence`	↑	308	767	1069	1170	3314
	↓	119	294	219	267	899
`rate_of_change`	↑	72	212	383	554	1221
	↓	6	38	45	45	134

structurally biased toward confidence-raising signals. Rather than indicating an extraction failure, this asymmetry reflects a direct imprint of editorial selection within the source corpus.

The failure mode for `fatigue` follows a different pattern. Six of the 12 reviewed `fatigue` cases were directionally incorrect, and all six involved injury-related articles. However, five other injury-related cases were handled correctly, resulting in an appropriate increase in `fatigue`. Rather than a systematic misreading of injury as rest, this variance indicates an inconsistency in how the same prompt instruction is applied across equivalent inputs. Because opposite adjustments were produced by the same type of evidence, it is likely that articles describing an absence in indirect terms, without explicit injury confirmation, were interpreted as precautionary rest or rotation. This misinterpretation ultimately pulled the `fatigue` score downward. While the prompt instruction mapping injury to high physical load is theoretically sound, its reliable application requires explicit cues in the input text. Therefore, upstream article clarity determines `fatigue` accuracy to a higher degree than prompt design alone.

The presence of incorrect or outdated information in 56.7% of the reviewed rows indicates that the primary source of error lies upstream of the reasoning step. While the phased article window (see 4.2.2) was implemented to mitigate the risk of stale context, analysis revealed that some of these errors actually stemmed from recent articles, published shortly before the match, that contained factually inaccurate reporting. Crucially, despite being supplied with this flawed evidence, the model produced highly coherent and logical outputs based strictly on the information available to it. Because the model faithfully applied the provided text to generate its adjustments without introducing hallucinations, these deviations represent failures in source data accuracy rather than

deficiencies in the model's reasoning capabilities.

While a positional concentration toward forwards was present in the audited sample, examining the full dataset (Table 4.11) shows that this concentration is not a matter of how often forwards are modified. Across the entire corpus, forwards received 3261 feature modifications and midfielders 3200, an effective tie, with defenders (2459) and goalkeepers (633) some way behind. Forwards therefore contribute roughly a third of all modifications yet make up two-thirds of the high-deviation sample, an over-representation that stems from the magnitude of their adjustments rather than their frequency. Because attackers typically generate highly polarised media narratives, such as scoring streaks or prominent goal droughts, the model produces larger absolute score adjustments for these players, causing them to dominate the top deviation percentiles.

Regarding magnitude calibration, the most common adjustment interval of 0.4 to 0.6 corresponds to the typical step size applied when the model encounters substantive, explicit textual evidence, such as a confirmed injury return or a standout match performance. Finally, the observed conservative bias, measured at 16.7% under-adjusted against only 6.7% over-adjusted, suggests that the model successfully hedges in ambiguous cases, preferring to defer to the Stats-Context prior rather than overclaiming when narrative support is weak.

Taken together, the audit evidence firmly answers **RQ1**(3.3.1) affirmatively. The LLM successfully extracts structurally coherent, aligned, and directionally valid signals from unstructured news articles. The model's adjustments moved in the correct conceptual direction in four out of five cases, backed by strong textual evidence and a conservative magnitude calibration that scales adjustments logically. Where directional misalignments did occur, most notably within the `fatigue` feature, they were driven by narrative ambiguity and a lack of explicit cues in the text, rather than flaws in the LLM's reasoning. Furthermore, while stale or inaccurate articles occasionally caused the system to output values that did not reflect the real-world match context, the LLM was able to extract information from the text accurately and without hallucination. This confirms that the model's core extraction capabilities are highly reliable, and that remaining errors are a function of upstream data clarity rather than LLM reasoning failures.

4.5 Predictive Performance

The five model configurations were evaluated independently for each of the three outfield positions on the chronological held-out test set described in Section 4.3.3.

4.5.1 Results

Test-set performance for the five model configurations across the three outfield positions is reported in Table 4.12, with RMSE taken as the primary criterion and MAE and MSE retained as secondary diagnostics. The per-bin RMSE breakdowns, computed over rating bins of width 0.1 across the [6.0, 8.0] target range, are shown in Figure 4.6, with the exact per-bin values reported in Tables D.1, D.2, and D.3 in Appendix D. Predicted-versus-actual scatter plots were produced for all fifteen position and variant combinations, each showing the test-set predictions against the

observed ratings together with the identity line and a band of ± 0.5 around it, and the full set is collected in Appendix D. These plots are referenced through this section wherever the dispersion of a given configuration bears on the interpretation of its aggregate score.

Among midfielders, the Full-Context variant recorded the lowest error on all three metrics, at an RMSE of 0.3837, an MAE of 0.3079, and an MSE of 0.1472, so on a typical match the model missed the observed rating by roughly three tenths of a point. This corresponds to a 3.4% reduction in RMSE relative to the Statistical Baseline at 0.3971 and 4.3% relative to the Average Baseline at 0.4010, and the permutation test (see, Section 4.3.6.2) confirms that both gains are statistically significant, at $p = 0.0004$ and $p = 0.0036$ respectively. No other contextual variant reproduced the gain. Stats-Context tracked the Statistical Baseline almost exactly at 0.4011, a difference indistinguishable from chance ($p = 0.38$), so re-encoding the statistical history into qualitative constructs added nothing on its own, and Article-Context produced the weakest midfielder scores of any configuration at an RMSE of 0.4430, significantly worse than both baselines ($p < 0.0001$). The per-bin breakdown in Table D.3 locates the Full-Context advantage in the four contiguous bins between 6.5 and 6.9, among the most populated in the range, with the baselines taking the dense bins immediately flanking this cluster, between 6.3 and 6.5 and again between 6.9 and 7.1, and the remaining, sparser bins split across the variants.

A different ordering emerged for defenders. The Average Baseline produced the lowest overall error, at an RMSE of 0.4331, an MAE of 0.3421, and an MSE of 0.1875, although its margin over the two strongest contextual variants is not statistically significant ($p = 0.18$ against Stats-Context and $p = 0.13$ against Full-Context), so the leading defender configurations are statistically indistinguishable from one another. The neural variants nonetheless separated sharply from their own statistical reference as the Statistical Baseline recorded the weakest result in the entire study at an RMSE of 0.4998, while appending the contextual vectors reduced that error by roughly 11% regardless of source, a gain the permutation test confirms for all three variants ($p < 0.0001$), leaving them within 0.0064 RMSE of one another (Stats-Context 0.4442, Full-Context 0.4460, Article-Context 0.4506). The per-bin results in Table D.2 divide the range into clear regions where the Average Baseline held the central bins between 6.5 and 7.2, the contextual variants led every bin below 6.5, and the Statistical Baseline, despite its weak aggregate, returned the lowest error in every bin above 7.2.

For forwards the Average Baseline was again the strongest configuration overall, at an RMSE of 0.3870, an MAE of 0.2953, and an MSE of 0.1497, with a wider margin over the neural variants than for defenders. The best neural variant was Article-Context at an RMSE of 0.4301, which stayed 11.1% above the Average Baseline while improving on the Statistical Baseline at 0.4621 by 6.9%, and Stats-Context gave a smaller 5.4% improvement over the same baseline at 0.4373. Both improvements over the Statistical Baseline are statistically significant ($p < 0.0001$ and $p = 0.0001$), as is the deficit of every neural configuration against the Average Baseline ($p < 0.0001$ in all four cases). Full-Context was the only contextual configuration in the study to fall below its own Statistical Baseline, producing the weakest neural forward scores at an RMSE of 0.4637, although this difference is statistically indistinguishable from the Statistical Baseline ($p = 0.83$).

The per-bin decomposition in Table D.1 splits the range almost cleanly. The Average Baseline held every bin between 6.0 and 7.0, while Stats-Context led the bins between 7.0 and 7.9, so the contextual variants retained an advantage for forwards only on above-average performances.

Two regularities recurred across the configurations, both visible in Figure 4.6. Per-bin RMSE was lowest in the central bins, near 0.20 to 0.35 between roughly 6.5 and 7.1, and rose toward both edges of the range, so each curve takes the broad form of a valley. The rise was asymmetric. Error climbed steeply toward the upper tail in all fifteen series, reaching between 0.79 and 1.19 in the 7.9 to 8.0 bin, while the climb toward the lower tail varied with the configuration. For the two baselines at every position, and for the contextual variants on midfielders and forwards, the lower extreme rose into the 0.65 to 0.93 range and produced a roughly symmetric U. For the three contextual variants on defenders the lower arm stayed much flatter, peaking near 0.52 to 0.54, since those variants returned the lowest error in every defender bin below 6.5, giving a right-skewed profile rather than a symmetric U.

The scatter plots in Appendix D account for the shared central concentration. In all fifteen plots the predictions collapse into a band roughly one rating point wide, concentrated between about 6.4 and 7.4, while the observed ratings span the full [6.0, 8.0] interval. High observed ratings are pulled downward and low observed ratings pushed upward, the signature of regression toward the training mean. The effect is clearest in the Average Baseline on midfielders (Figure D.2) and recurs across the remaining plots with only second-order differences in spread, the tightest clouds belonging to the Average variants and to Full-Context on midfielders. A few configurations add a vertical offset to this compression, most visibly the Statistical Baseline on defenders (Figure D.4), whose cloud sits above the identity line in keeping with its last-place aggregate score. No configuration produced predictions that follow the identity line into either tail.

4.5.2 Analysis

The three contextual variants were built so that the pairwise contrasts isolate the contribution of each information source, and the results become legible only when read against that design. Stats-Context measures what is gained by re-encoding the existing statistical history into the seven bounded qualitative constructs, with no media input at all. Article-Context measures what editorial text adds when applied to that prior without any statistical check. Full-Context measures what media evidence contributes once it has been cross-referenced against the statistical record before any override is accepted. Each position orders these three variants differently, and each ordering points to a different mechanism.

For midfielders the ordering is the most informative of the three. The gain over both baselines belongs to Full-Context alone, and the per-bin table shows that it is earned on the densest central bins rather than on a handful of well-placed outliers, so the model predicts the typical midfielder performance more accurately than either benchmark. The remaining variants explain where that gain originates. Stats-Context matches the Statistical Baseline, so the qualitative re-encoding contributes nothing on its own, while adding the unverified article layer on top of that prior raises midfielder error from 0.4011 to 0.4430. This is the behaviour an unverified article channel would

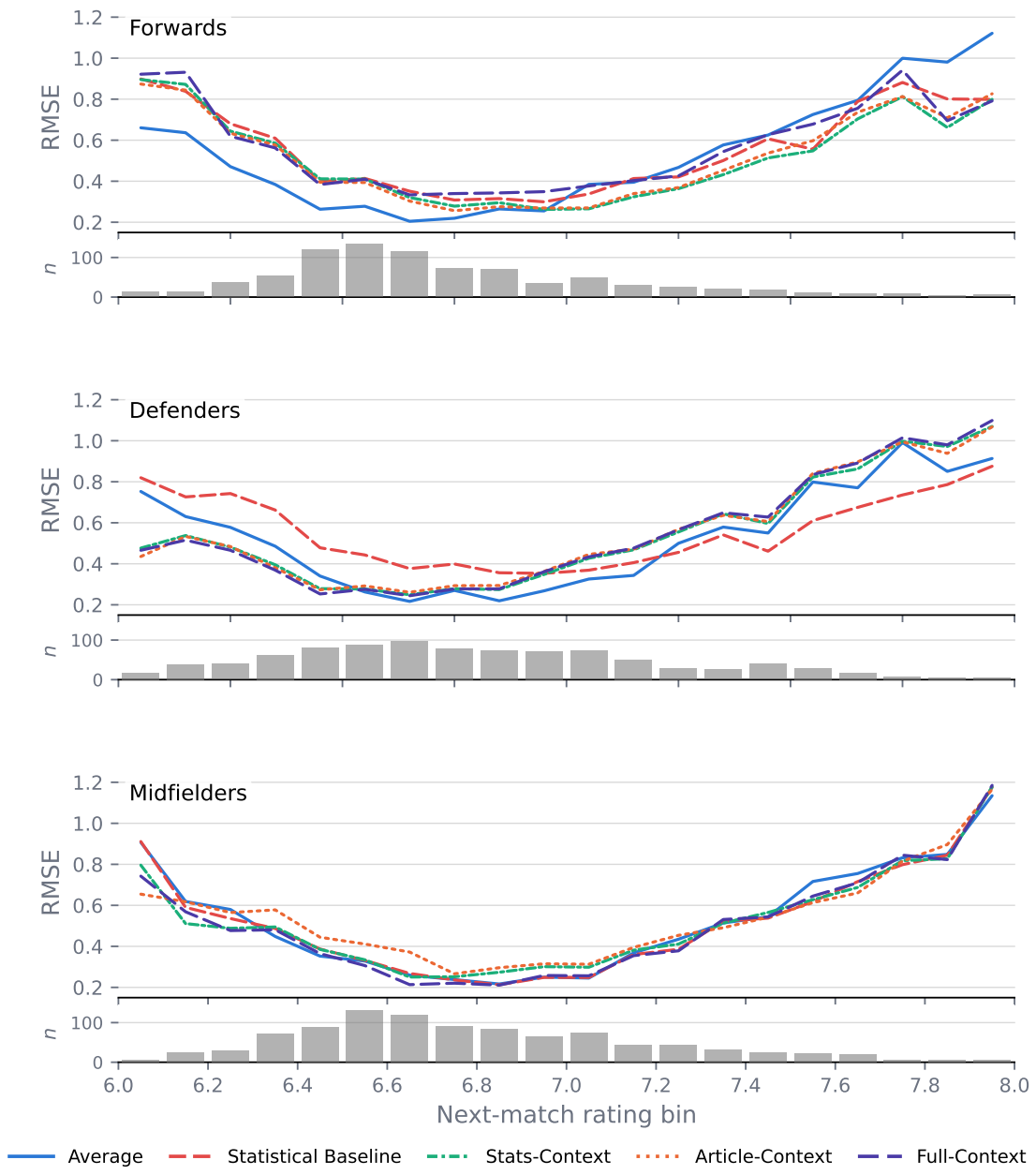

Figure 4.6: Per-bin test-set RMSE for each model configuration across the $[6.0, 8.0]$ next-match rating range, by position. Each curve is a model configuration; the thin strip beneath each panel reports the number of test rows n per bin. Exact values appear in Tables D.1, D.2 and D.3 in Appendix D.

Table 4.12: Test-set metrics per position and variant (MAE, RMSE, MSE). Bold marks the best result for each metric within a position.

Position	Variant	MAE	RMSE	MSE
MF	Average Baseline	0.3120	0.4010	0.1608
	Statistical Baseline	0.3165	0.3971	0.1577
	Stats-Context	0.3159	0.4011	0.1609
	Article-Context	0.3475	0.4430	0.1963
	Full-Context	**0.3079**	**0.3837**	**0.1472**
DF	Average Baseline	**0.3421**	**0.4331**	**0.1875**
	Statistical Baseline	0.4037	0.4998	0.2498
	Stats-Context	0.3495	0.4442	0.1973
	Article-Context	0.3560	0.4506	0.2031
	Full-Context	0.3497	0.4460	0.1989
FW	Average Baseline	**0.2953**	**0.3870**	**0.1497**
	Statistical Baseline	0.3595	0.4621	0.2135
	Stats-Context	0.3522	0.4373	0.1913
	Article-Context	0.3445	0.4301	0.1850
	Full-Context	0.3615	0.4637	0.2150

be expected to produce, since the reliability audit in Section 4.4 attributes the errors to upstream data quality rather than to model reasoning, finding stale or inaccurate reporting in 56.7% of the cases it examined, the kind of noise a pipeline with no statistical check cannot remove. The signal in the article corpus becomes net positive only once it is validated against match history, as it is under Full-Context, and the permutation test quantifies both sides of this contrast. The unverified article layer degrades midfielder error significantly ($p < 0.0001$) while the verified layer improves it significantly ($p = 0.0004$), so the verification step is the difference between a harmful and a beneficial signal rather than a marginal refinement. Editorial media does therefore carry predictive information about midfielder performance, but that information is recoverable only through a verification step that filters narrative claims against the statistical record, and the error reduction sought by the second research question is realised for this position, and for this position alone, under exactly that condition.

For defenders and forwards the same question has to be answered in two parts, because the two relevant comparisons disagree. Measured against the Statistical Baseline, the contextual features reduced error for both positions in their best configuration, and for defenders the reduction of about 11% in RMSE was the largest in the study, independent of which source generated the vectors. Measured against the Average Baseline, no neural configuration overtook it for either position, with the best defender variant 2.6% above the Average Baseline, a margin not statistically significant ($p = 0.18$), and the best forward variant 10% above it, a deficit the permutation test confirms as significant ($p < 0.0001$). The strength of that benchmark is itself a result worth reading. Player

ratings in this dataset are strongly autocorrelated at the individual level, so a player-specific recent mean is a demanding target, consistent with the recurring finding that lagged individual output is among the strongest predictors available in player forecasting [39, 60]. The two error measures separate the baselines for midfielders in a way that makes this concrete, since the Average Baseline posts the lower MAE, at 0.3120 against 0.3165, but the higher RMSE, at 0.4010 against 0.3971, which places the strength of the per-player mean in the typical match and its weakness in the volatile results that RMSE penalises most heavily. To beat it, a neural model has to encode that recent level internally and additionally learn when to depart from it, and the departures concentrate in the high-variance, event-driven outcomes that pre-match features constrain only weakly. The contextual vectors did add information beyond the statistical features on their own, but for two of the three positions the resulting models still could not recover enough of that deviation signal to overtake a benchmark that simply assumes performance persists.

The defender model ranking localises the contextual gain differently from the midfielder case. The three variants sit within 0.0064 RMSE of one another and the article-free Stats-Context variant is marginally the best of them, so the 11% improvement cannot be credited to media content because it is fully present without any article input. What the three share is the seven-construct encoding, which suggests that the bounded qualitative summary works as a useful low-dimensional compression of conditions that the raw defender feature space represents poorly. This fits the statistical profile of the position reported in the literature, where defensive output depends on collective team structure and volume rather than on individually attributable events [36, 54]. A compact representation of role continuity, fixture stakes, and opponent strength plausibly captures most of what is predictable about a defender's next rating, while the granular per-match action counts add more noise than signal, which the last-place finish of the Statistical Baseline makes visible. The same encoding also accounts for the asymmetry in the defender error curves. The three contextual variants return the lowest error in every bin below 6.5, where the Average and Statistical Baselines perform poorly. Because the contextual variants already handle the low end well, their per-bin error rises only gently at the bottom of the range. Instead, their error rises only as ratings move toward the high tail, producing the right-skewed profile visible in the defender panel of Figure 4.6.

The forward ordering inverts the midfielder result, with Article-Context the best neural variant and Full-Context the worst, the latter posting an RMSE 0.3% higher than even the Statistical Baseline. Two mechanisms plausibly combine to produce this reversal. Forwards attract the most polarised and event-driven coverage of any role, which the audit reflects directly, since they account for 66.7% of the thirty highest-deviation adjustments while contributing only about a third of all contextual modifications, 3261 of 9553 in Table 4.11. Polarised coverage produces large contextual swings, and Full-Context, which modifies 95.9% of rows against 87.1% for Article-Context, propagates those swings most aggressively. The cross-referencing step that helps midfielders also depends on the statistical record being a stable anchor against which article claims can be tested, and forward statistics, dominated by low-frequency scoring events, are the most volatile of the three positions. Validating a narrative claim against an erratic record risks confirming noise with

noise, so where the anchor is stable the procedure filters the corpus and where it is itself unstable the same procedure can amplify spurious agreement between an inflated story and a short statistical streak. This mechanism is inferred from the variant ordering rather than measured directly, and Section 5.4 returns to it as a hypothesis for future testing. The per-bin structure adds one qualification, since the contextual variants led every forward bin above 7.0, which indicates that signals such as confidence and short-term momentum do carry information about the chance of an above-baseline display, even where the typical sub-7.0 performance is best predicted by simple persistence.

Beneath these position-specific orderings, the central concentration of accuracy and the compression visible across all fifteen scatter plots show that every neural variant learned a conservative mapping that regresses toward the centre of the training distribution. The per-bin curves in Figure 4.6 make the same point, with error lowest in the central bins and rising toward the tails, steeply toward the high tail for every configuration and toward the low tail for all but the three defender contextual variants, which alone hold the lower bins down. The inverse-frequency sample weighting described in Section 4.3.4 reduced this behaviour without removing it, which is unsurprising given that the extreme bins hold between 5 and 29 test rows per position, too few to reward tail-sensitive parameters during validation-based selection. The contextual gains documented above are therefore redistributions of accuracy within the predictable core of the distribution, roughly between 6.5 and 7.5, while exceptional performances at the tails remain mispredicted by margins between about 0.5 and 1.2 RMSE depending on the tail and configuration. The MSE column in Table 4.12 adds no independent evidence here, because it is the square of RMSE and orders the configurations identically. It is retained for completeness, and the larger gains it reports, such as the 6.7% midfielder improvement against 3.4% in RMSE, follow from the squared scale rather than from any separate effect.

Chapter 5

Conclusion

This chapter draws the thesis to a close. It revisits the three research questions against the evidence gathered in Chapter 4 and reads that evidence as a single body of findings rather than a set of per-position results, drawing out what the study established and where its claims stop. The limitations that bound those conclusions are then set out, followed by the threats that qualify their validity, and the chapter ends with the work that would carry the framework into new settings and test directly the mechanisms that could here only be inferred.

5.1 Discussion

The evaluation in Chapter 4 answered the three research questions, and the answers are read here at the level of the study as a whole. The per-position mechanisms behind them are set out in the analysis of Section 4.5.2 and are not repeated.

At the aggregate level the headline is narrow but statistically firm. The only configuration to improve on both baselines at once was the Full-Context model for midfielders, which lowered RMSE by 3.4% against the Statistical Baseline and 4.3% against the Average Baseline, which are bothstatistically significant ($p = 0.0004$ and $p = 0.0036$). For defenders and forwards no neural configuration beat the Average Baseline. For forwards that deficit is itself significant ($p < 0.0001$), while for defenders the best variants trail by margins indistinguishable from chance ($p \geq 0.13$), leaving the leading defender configurations statistically tied with the persistence benchmark. The practical predictive gain from the contextual pipeline is therefore confined to a single position.

The first research question asked whether a large language model can extract structurally coherent, aligned, and directionally valid contextual signals from unstructured football news. The reliability audit in Section 4.4 answered this affirmatively, reaching 80.0% directional accuracy and a mean row score of 0.84 out of 1.0 across the thirty highest-deviation cases, with confidence, role stability, and rate of change recovered with full directional agreement. The signal is recoverable, with the qualification that its reliability is uneven across constructs and weakest where the article corpus carries stale reporting.

The second research question asked to what extent these signals reduce next-match rating error. The answer is conditional. Validated contextual signals lowered error against both baselines for one configuration and one position, the midfielder Full-Context model, and both reductions are statistically significant. The same test shows the unverified Article-Context variant significantly increasing midfielder error relative to both baselines ($p < 0.0001$), so verification is not an optional refinement but the condition under which the media signal changes sign. For defenders the contextual encoding reduced error against the Statistical Baseline by about 11% ($p < 0.0001$), but that gain was realised without any article input and so cannot be attributed to media content, and the defender models remained 2.6% above the Average Baseline, a margin not statistically significant. For forwards no contextual configuration matched the Average Baseline ($p < 0.0001$). The media signal is therefore net positive only where editorial evidence is cross-checked against the statistical record and only for a role whose performance responds to it, for reasons developed in Section 4.5.2.

The third research question asked whether the benefit varies across positions and which roles are most sensitive. It does vary, in a clear ordering from midfielders through defenders to forwards, which tracks how far each role's rating is built from continuous, individually attributable involvement rather than from discrete and largely stochastic events. The mechanism behind that ordering, including the mean-stability of defenders and the noise-amplification risk for forwards under cross-referencing, is set out in Section 4.5.2. The variation is a property of the positions themselves rather than of how much coverage the press supplies, since midfielders and forwards receive comparable modification volumes.

Overall, the three answers converge on a qualified confirmation of the central hypothesis. The claim that LLM-extracted contextual signals from editorial media reduce next-match rating prediction error relative to purely statistical models holds for midfielders under the Full-Context configuration, where media evidence verified against statistical history improved all three metrics across the most populated regions of the rating distribution, with both improvements confirmed as statistically significant by the permutation test. It does not hold for defenders, whose gains arise from the qualitative encoding rather than from media content, nor for forwards, where no contextual configuration reached the Average Baseline. The hypothesis is therefore supported in a restricted rather than general form. Editorial media carries recoverable predictive signal for roles whose performance reflects continuous individual involvement, and that signal becomes net positive only when its extraction includes a verification step that cross-checks narrative claims against the statistical record.

5.2 Limitations

The evaluation rests on a single competition. All data was drawn from Liga Portugal across the 2024/2025 season and the first twenty-seven matchdays of 2025/2026, so the league's particular distribution of playing styles, scoreline patterns, and editorial conventions is baked into every result. Whether the midfielder gain transfers to leagues with different tactical profiles or denser,

differently toned press coverage cannot be established from this data, and the positional ordering may shift where the relationship between role and rating differs.

The contextual signals were produced by a single language model. Every vector was generated with Claude Opus 4.7 under one prompt schema, so the study measures what one extraction setup recovers rather than the capability of large language models in general. The findings should be read as a lower bound tied to that setup, and the fatigue construct in particular, which the audit found least reliable, might respond to a more constrained prompt design.

The target range was narrowed to ratings between 6.0 and 8.0. Ratings below 6.0 follow from structurally unforeseeable events such as red cards and decisive errors, and ratings above 8.0 are too sparse to learn from, so the restriction was a deliberate scoping choice. The consequence is that the framework is evaluated only on the central body of outcomes and not on the standout or breakout matches that fall in the excluded tails. Those tail matches are precisely what the talent identification and undervalued-player recruitment set out in the motivation of Section 1.2, together with the fantasy football use case introduced in Section 1.1, most need to anticipate, so the capability the work was framed around in Chapter 1 is the one this scoping choice defers. A route back toward that capability is developed among the future directions in Section 5.4.

The dataset spans two seasons. The chronological split leaves 845 to 986 test rows per position, and the extreme rating bins contain as few as five rows, which constrains the confidence that can be placed in any per-bin claim near the edges of the range.

5.3 Threats to Validity

The most consequential threat concerns how the rating truncation described above interacts with the sequence model. Rows outside [6.0, 8.0] were removed before the per-player sequences were assembled, so the five previous matches consumed by the LSTM are the five most recent surviving matches rather than the five most recent in fact. Two distortions follow. The recent history loses exactly the high-variance results that most strongly signal an impending regression or breakout, so the model is trained on an artificially smoothed view of form. The surviving matches are also no longer temporally contiguous, which means the uniform spacing the LSTM implicitly assumes between sequence steps does not hold. Both effects push the model toward the conservative, mean-reverting mapping documented across the fifteen scatter plots, and because the persistence baselines are computed on the same surviving rows, the truncation may handicap the neural variants more than the benchmarks they are compared against. The size of this effect was not measured and should be treated as a plausible contributor to the tail failure rather than a quantified cause.

A second threat concerns construct validity in the audit. Directional accuracy was assessed on the thirty highest-deviation cases, a sample deliberately weighted toward large adjustments and toward forwards in particular. The 80.0% figure therefore characterises the model's behaviour where it intervenes most aggressively, which is the right place to look for failure, but it should not be read as an accuracy estimate over the full corpus, most of which involves smaller or no adjustments.

A third threat is the dependence of the cross-referencing mechanism on the statistical record being a stable anchor. The interpretation offered for the forward result, that validating a narrative claim against a volatile scoring record can confirm noise with noise, is inferred from the variant ordering rather than measured directly. It is consistent with the evidence but not established by it, and an alternative account in which the forward articles are simply less informative cannot be excluded with the present data.

5.4 Future Work

The most immediate extension is to repeat the evaluation on other competitions. A multi-league corpus would test whether the midfielder gain and the positional ordering hold beyond Liga Portugal and would enlarge the sparsely populated rating bins that limited tail-sensitive model selection in the present study.

A second strand concerns the extraction step. Comparing several models and several prompt designs would show how much of the signal quality is attributable to the extractor and how much to the source text. The fatigue construct is the natural starting point, since its errors were traced to narrative ambiguity rather than reasoning failure, and a more constrained prompt that requires explicit injury or workload cues before adjusting the score would test whether its accuracy can be raised without changing the rest of the schema.

The sequence-truncation threat identified in Section 5.3 invites two complementary remedies. The first is to quantify the effect by measuring how many sequences contained at least one removed match and how far their predictions differ from intact sequences. The second is to remove the cause by keeping the full rating history in the input window and restricting only the prediction target to the studied range, so that the model predicts central outcomes from an unbroken and correctly spaced view of recent form. This second remedy repairs the integrity of the input sequence but leaves the prediction target bounded at 8.0, so on its own it sharpens central-range accuracy without restoring the ability to predict the exceptional matches the truncation removed.

Recovering that ability is the aim of a further line of work that targets the tail performances directly, and it is this strand, rather than the input-history repair above, that reopens the talent identification, undervalued-player recruitment, and fantasy applications the study was framed around in Chapter 1. Approaches worth examining include a two-stage formulation that first classifies a match as typical or exceptional and then predicts within the identified regime, asymmetric or quantile losses that reward correct tail predictions more heavily than the inverse-frequency weighting used here, and the larger tail samples that a multi-league corpus would provide. For this tail work to be sound the input-history repair is its necessary companion, since predicting the tails from sequences that are themselves missing the high-variance matches would undermine the attempt from the start.

Finally, the forward cross-referencing mechanism proposed in Chapter 4 predicts a specific relationship between a position's statistical volatility and the benefit of the verification step. Reporting the variance of rating sequences per position, and ablating the cross-referencing step sepa-

rately for high-volatility and low-volatility players, would show whether the forward degradation is driven by the instability of the anchor or by the articles themselves being less informative.

References

[1] A. Al-Bustami and Z. Ghazal. "From Players to Champions: A Generalizable Machine Learning Approach for Match Outcome Prediction with Insights From the FIFA World Cup". In: *2025 IEEE International Conference on Electro Information Technology (eIT)*. 2025 IEEE International Conference on Electro Information Technology (eIT). May 29–31, 2025. DOI: 10.1109/eIT64391.2025.11103598.

[2] A. Agarwal and D. Toshniwal. "Application of Lexicon Based Approach in Sentiment Analysis for Short Tweets". In: 2018. DOI: 10.1109/ICACCE.2018.8441696.

[3] Aisha Alansari and Hamzah Luqman. "Large Language Models Hallucination: A Comprehensive Survey". In: arXiv:2510.06265 (Mar. 2026). arXiv:2510.06265 [cs.CL]. DOI: 10.48550/arXiv.2510.06265.

[4] Tomás Duarte de Almeida. "Predicting sports events using Twitter sentiment analysis and neural networks". en. In: ().

[5] Stefan Altmann et al. "Match-related physical performance in professional soccer: Position or player specific?" In: *PLoS ONE* 16.9 (Sept. 2021), e0256695. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0256695.

[6] S. Aruna et al. "Machine Learning Driven Predictions for IPL Match Results". In: 2025. DOI: 10.1109/ICPCSN65854.2025.11034925.

[7] Diêgo Augusto et al. "Contextual variables affect peak running performance in elite soccer players: A brief report". In: *Frontiers in Sports and Active Living* 4 (Sept. 2022), p. 966146. ISSN: 2624-9367. DOI: 10.3389/fspor.2022.966146.

[8] Robert Bajons. "Evaluating Player Performances in Football: A Debiased Machine Learning Approach". In: *Proceedings of the 2023 6th International Conference on Mathematics and Statistics*. ICoMS '23. New York, NY, USA: Association for Computing Machinery, Dec. 2023, pp. 135–140. ISBN: 979-8-4007-0018-7. DOI: 10.1145/3613347.3613368.

[9] Y. Bengio, P. Frasconi, and P. Simard. "The problem of learning long-term dependencies in recurrent networks". In: *IEEE International Conference on Neural Networks*. Mar. 1993, 1183–1188 vol.3. DOI: 10.1109/ICNN.1993.298725.

[10] Mark Carey. *How the growth of wearable technology is transforming football*. Oct. 2023. URL: https://www.nytimes.com/athletic/4966509/2023/10/19/wearable-technology-in-football/.

[11] Zhong An Chen. "TA Hybrid Machine Learning Framework for Soccer Match Outcome Prediction: Incorporating Bivariate Poisson Distribution". In: *ITM Web Conf.* (Jan. 2025). ISSN: 2271-2097. DOI: 10.1051/itmconf/20257003020.

[12] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation". en. In: *BioData Mining* 14.1 (Feb. 2021), p. 13. ISSN: 1756-0381. DOI: 10.1186/s13040-021-00244-z.

[13] Felipe Dambroz, Filipe Manuel Clemente, and Israel Teoldo. "The effect of physical fatigue on the performance of soccer players: A systematic review". en. In: *PLOS ONE* 17.7 (July 2022), e0270099. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0270099.

[14] Deloitte. *Deloitte Football Money League 2025*. Jan. 2025. URL: https://www.deloitte.com/br/pt/Industries/tmt/analysis/deloitte-football-money-league.html.

[15] M. Elsharkawi, R.H. Ali, and T.A. Ali Khan. "Crafting a Player Impact Metric through Analysis of Football Match Event Data". In: *Journal of Computational Mathematics and Data Science* 15 (2025). DOI: 10.1016/j.jcmds.2025.100115.

[16] fbref. URL: https://fbref.com/en/.

[17] FIFA. *World Cup 2022 in Numbers*. 2022. URL: https://inside.fifa.com/tournament-organisation/world-cup-2022-in-numbers.

[18] Leon Forcher et al. "Does Technical Match Performance in Professional Soccer Depend on the Positional Role or the Individuality of the Player?" English. In: *Frontiers in Psychology* 13 (May 2022). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2022.813206. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.813206/full.

[19] Fundação para a Ciência e a Tecnologia (FCT) – FCCN. *IAedu: Inteligência Artificial para o ensino superior e investigação*. https://iaedu.pt/pt. 2025.

[20] Jianxiong Gao, Yi Cheng, and Jianwei Gao. "Predicting sport event outcomes using deep learning". en. In: *PeerJ Computer Science* 11 (July 2025), e3011. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.3011.

[21] Romain Gauriot. "Fooled by Performance Randomness: Overrewarding Luck". In: *The Review of Economics and Statistics* (Oct. 2019). ISSN: 0034-6535. DOI: 10.1162/rest_a_00783.

[22] Luiz Octávio Gavião et al. *Evaluation of soccer players under the Moneyball concept*. First. Routledge, 2023. ISBN: 9781003375968.

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. en. Google-Books-ID: omivDQAAQBAJ. MIT Press, Nov. 2016. ISBN: 978-0-262-33737-3.

[24] Peter Hassard and Dermot Kerr. "Predicting Football Match Outcomes Using Event Data and Machine Learning Algorithms". en. In: *2024 35th Irish Signals and Systems Conference (ISSC)*. Belfast, United Kingdom: IEEE, June 2024, pp. 1–6. ISBN: 979-8-3503-5298-6. DOI: 10.1109/ISSC61953.2024.10603147. URL: https://ieeexplore.ieee.org/document/10603147/.

[25] Benjamin Hendricks. "Sports Analytics with Natural Language Processing: Using Crowd Sentiment to Help Pick Winners in Fantasy Football". en. In: (May 2022).

[26] Francisco Henriques. "HOW CAN BIG DATA HELP FOOTBALL CLUBS ACHIEVE COMPETITIVE ADVANTAGE". In: (Mar. 2018).

[27] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735.

[28] I. Hosein, K. Baboolal, and P. Hosein. "Models for Predicting the Success of a Footballer". In: *2024 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*. journalAbbreviation: 2024 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD). Nov. 2024, pp. 1–6. DOI: 10.1109/ICTMOD63116.2024.10959128.

[29] Yebowen Hu et al. "SportsMetrics: Blending Text and Numerical Data to Understand Information Fusion in LLMs". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 267–278. DOI: 10.18653/v1/2024.acl-long.17.

[30] Ziwei Ji et al. "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55.12 (Dec. 2023). arXiv:2202.03629 [cs.CL], pp. 1–38. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3571730.

[31] Jacky Jiang and Jerry Cai. "GoalNet: Unveiling-Hidden-Pivotal-Players-A-GNN-Based-Soccer-Player- Evaluation-System". en. In: ().

[32] Jonathan D Kosy John S Theodoropoulos Jeremy Bettle. "The use of GPS and inertial devices for player monitoring in team sports: A review of current and future applications". In: *Orthopedic Reviews* (Apr. 2020). DOI: 10.4081/or.2020.7863.

[33] M.A. Khan et al. "Analysis of the Smart Player's Impact on the Success of a Team Empowered with Machine Learning". In: *Computers, Materials and Continua* 66.1 (2021). DOI: 10.32604/cmc.2020.012542.

[34] Masato Kikuchi et al. "Confidence Interval of Probability Estimator of Laplace Smoothing". In: *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*. arXiv:1709.08314 [cs.IT]. Aug. 2015, pp. 1–6. DOI: 10.1109/ICAICTA.2015.7335387. URL: http://arxiv.org/abs/1709.08314.

[35] Carlos Lago-Peñas. "The Role of Situational Variables in Analysing Physical Performance in Soccer". In: *J Hum Kinet* (Dec. 2012). ISSN: 1640-5544. DOI: 10.2478/v10078-012-0082-9.

[36] D. Malikov and J. Kim. "Beyond xG: A Dual Prediction Model for Analyzing Player Performance Through Expected and Actual Goals in European Soccer Leagues". In: *Applied Sciences (Switzerland)* 14.22 (2024). DOI: 10.3390/app142210390.

[37] Farzam Manafzadeh, Changiz Eslahchi, and Hadi Nobari. "Success Score: A Deep Learning Framework for Predicting Football Match Outcomes and Evaluating Team Performance". en. In: (Oct. 2025). DOI: 10.21203/rs.3.rs-7736577/v1. URL: https://www.researchsquare.com/article/rs-7736577/v1.

[38] S. Manish, V. Bhagat, and R.M. Pramila. "Prediction of Football Players Performance Using Machine Learning and Deep Learning Algorithms". In: 2021. DOI: 10.1109/INCET51464.2021.9456424.

[39] C. Markopoulou, G. Papageorgiou, and C. Tjortjis. "Diverse Machine Learning for Forecasting Goal-Scoring Likelihood in Elite Football Leagues". In: *Machine Learning and Knowledge Extraction* 6.3 (2024). DOI: 10.3390/make6030086.

[40] Oliver Müller Matthew Caron. "TacticalGPT: Uncovering the Potential of LLMs for Predicting Tactical Decisions in Professional Football". In: (2023).

[41] Leah M Monsees. ""There is a lot more potential" - practitioner perspectives on technology and data-driven talent identification, selection, and development in a German Bundesliga academy". In: *International Journal of Sports Science and Coaching* (Jan. 2025). ISSN: 1747-9541. DOI: 10.1177/17479541241308519.

[42] opta. URL: https://www.statsperform.com/opta/.

[43] Ioannis Antoniadis Paraskevas Chatziparaskevas Vaggelis Saprikis. "The impact of information systems and data science on management in modern professional football: Moneyball theory and the development model of Brentford FC". In: *AIP Conf. Proc.* (Oct. 2024). ISSN: 0094-243X. DOI: 10.1063/5.0237053.

[44] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks". In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML'13. Atlanta, GA, USA: JMLR.org, June 2013, pp. III-1310-III–1318.

[45] Weichen Peng. "Machine Learning Models for Nba Game Prediction". en. In: *ITM Web of Conferences* 78 (2025). Ed. by K. Subramaniya, p. 01011. ISSN: 2271-2097. DOI: 10.1051/itmconf/20257801011.

[46] Paul Pettitt. *The Importance of Accurate and Reliable Sports Data*. URL: https://www.statsperform.com/resource/the-importance-of-accurate-and-reliable-sports-data-in-todays-sports-industry/.

[47] Dhanesh Ramachandram and Graham W. Taylor. "Deep Multimodal Learning: A Survey on Recent Advances and Trends". In: *IEEE Signal Processing Magazine* 34.6 (Nov. 2017), pp. 96–108. ISSN: 1558-0792. DOI: 10.1109/MSP.2017.2738401.

[48] K. Ravichandran et al. "A Novel Graph Based Approach to Predict Man of the Match for Cricket". In: Communications in Computer and Information Science. Vol. 1241 CCIS. 2020. DOI: 10.1007/978-981-15-6318-8_48.

[49] Zia Ur Rehman et al. "Predict the Match Outcome in Cricket Matches Using Machine Learning". In: VOLUME 03 (May 2022), pp. 206–2012.

[50] Fátima Rodrigues and Ângelo Pinto. "Prediction of football match results with Machine Learning". In: *Procedia Computer Science*. International Conference on Industry Sciences and Computer Science Innovation 204 (Jan. 2022), pp. 463–470. ISSN: 1877-0509. DOI: 10.1016/j.procs.2022.08.057.

[51] Helen Ruud and Julius Jooste. "Investigating the interconnectedness of athletes' performance beliefs, coping ability, and mental health, and the efficacy of a REBT-inspired mobile app intervention". en. In: *Performance Enhancement & Health* 13.4 (Oct. 2025), p. 100369. ISSN: 22112669. DOI: 10.1016/j.peh.2025.100369.

[52] S. Said et al. "Soccer Events Summarization by Using Sentiment Analysis". In: 2016. DOI: 10.1109/CSCI.2015.59.

[53] Robert P. Schumaker and Osama K. Solieman. *Sports Data Mining: The Field*. Springer US, 2010. ISBN: 978-1-4419-6730-5.

[54] T. Sharma et al. "Comparative Analysis of Machine Learning Models for Predicting Top Goal Scorer and Goalkeeper Performance in Football". In: 2025. DOI: 10.1109/ICDT63985.2025.10986354.

[55] Q. Song and R. Lu. "Analysis of Sports Event Data Based on the LSTM". In: 2024. DOI: 10.1145/3675249.3675348.

[56] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. "Multimodal deep learning for biomedical data fusion: a review". In: *Briefings in Bioinformatics* 23.2 (Jan. 2022), bbab569. ISSN: 1467-5463. DOI: 10.1093/bib/bbab569.

[57] T. Pawar, P. Kalra, and D. Mehrotra. "Analysis of Sentiments for Sports Data Using RapidMiner". In: *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*. 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT). Aug. 16–18, 2018. DOI: 10.1109/ICGCIoT.2018.8752989.

[58] M. Tamimi and T. Tran. "Players' Performance Prediction for Fantasy Premier League, Using Transformer-based Sentiment Analysis on News and Statistical Data". In: *International Journal of Computer Science in Sport* 24.1 (2025). DOI: 10.2478/ijcss-2025-0008.

[59] E. Tiwari, P. Sardar, and S. Jain. "Football Match Result Prediction Using Neural Networks and Deep Learning". In: 2020. DOI: 10.1109/ICRITO48877.2020.9197811.

[60] K. van Arem, F. Goes-Smit, and J. Söhl. "Forecasting the Future Development in Quality and Value of Professional Football Players". In: *Applied Sciences (Switzerland)* 15.16 (2025). DOI: 10.3390/app15168916.

[61] W. Liu et al. "An Improved Evaluation Method for Soccer Player Performance Using Affective Computing". In: *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*. 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD). May 28–31, 2020. DOI: 10.1109/ICAIBD49809.2020.9137435.

[62] T. Wang and J. Wang. "Footballnet: A Deep Learning Network for Football Competition Prediction". In: 2025. DOI: 10.1145/3723936.3723973.

[63] Jason Wei et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". en. In: arXiv:2201.11903 (Jan. 2023). arXiv:2201.11903 [cs]. DOI: 10.48550/arXiv.2201.11903.

[64] V. Vanessa Wergin et al. "When Suddenly Nothing Works Anymore Within a Team – Causes of Collective Sport Team Collapse". In: *Front Psychol* (Nov. 2018). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.02115.

[65] zerozero. URL: https://www.zerozero.pt/.

Appendix A

Feature Sets

A.1 Raw Features

A.1.1 Player Match Statistics

Table A.1 lists the raw statistical columns prior to any feature engineering.

Table A.1: Raw feature set prior to engineering

Feature	Feature
minutes_played	goals
assists	yellow_cards
red_cards	passes_successful
passes_total	rating
shots_total	fouls_committed
tackles_total	dribbles_total
dribbles_successful	duels_total
duels_won	aerials_won
dispossessed	xg_total
aerials_total	steals
fouls_won	key_passes
dribbles_suffered	big_chances_created
captaincy	crosses_total

Continued on next page

Table A.1 (continued)

Feature	Feature
interceptions	shot_on_target
xg_on_target	big_chance_missed
blocks_total	hit_woodwork
crosses_successful	gk_saves_total
penalties_won	offsides
gk_punches	penalties_conceded
called_up	clearances_total
ball_recoveries	long_balls_total
long_balls_successful	touches
gk_high_claims	possession_lost
passes_in_final_third	gk_saves_inside_box
backward_passes	tackles_won
duels_lost	chances_created
shots_off_target	blocks_defensive
shots_blocked	turnovers
through_balls	through_balls_successful
offsides_provoked	

A.1.2 Fixtures

Table A.2 lists the columns present in the fixtures dataset.

Table A.2: Raw fixtures dataset columns

Feature	Feature
team_id	team_name
match_id	opp_id
opponent_name	match_date

Continued on next page

Table A.2 (continued)

Feature	Feature
is_home	season_id
team_points	team_league_pos
opp_points	opp_league_pos

A.1.3 Squads

Table A.3 lists the columns present in the squad compositions dataset.

Table A.3: Raw squad compositions dataset columns

Feature	Feature
match_id	team_id
player_id	player_name
position	

A.1.4 Articles

Table A.4 lists the columns present in the news articles dataset.

Table A.4: Raw news articles dataset columns

Feature	Feature
match_id	team_id
published_date	text

A.2 Engineered Features

A.2.1 Efficiency Ratio Features

Table A.5 lists the efficiency ratio features and indicates whether Laplace smoothing was applied to each.

Table A.5: Efficiency ratio features and whether Laplace smoothing was applied

Feature	Smoothed
pass_accuracy	Yes
dribble_success_rate	Yes
tackle_win_rate	Yes
aerial_win_rate	Yes
ground_duel_win_rate	Yes
long_ball_accuracy	Yes
cross_accuracy	Yes
shot_accuracy	Yes
through_ball_accuracy	Yes
key_pass_rate	Yes
big_chance_miss_rate	Yes
xg_per_shot	No
xg_conversion_rate	No

A.2.2 Per-90 Features

Table A.6 lists the per-90 features and the source column from which each was derived.

Table A.6: Per-90 features and their corresponding source columns

Feature	Source Column
goals_p90	goals
assists_p90	assists
passes_successful_p90	passes_successful
passes_in_final_third_p90	passes_in_final_third
dribbles_successful_p90	dribbles_successful

Continued on next page

Table A.6 (continued)

Feature	Source Column
tackles_won_p90	tackles_won
aerials_won_p90	aerials_won
duels_won_p90	duels_won
long_balls_successful_p90	long_balls_successful
crosses_successful_p90	crosses_successful
through_balls_successful_p90	through_balls_successful
shots_total_p90	shots_total
xg_total_p90	xg_total
xg_on_target_p90	xg_on_target
key_passes_p90	key_passes
chances_created_p90	chances_created
big_chances_created_p90	big_chances_created
touches_p90	touches
fouls_won_p90	fouls_won
fouls_committed_p90	fouls_committed
dribbles_suffered_p90	dribbles_suffered
offsides_provoked_p90	offsides_provoked
ball_recoveries_p90	ball_recoveries
clearances_total_p90	clearances_total
interceptions_p90	interceptions
steals_p90	steals
blocks_defensive_p90	blocks_defensive
shots_blocked_p90	shots_blocked
possession_lost_p90	possession_lost
gk_saves_total_p90	gk_saves_total
gk_saves_inside_box_p90	gk_saves_inside_box
gk_high_claims_p90	gk_high_claims

Continued on next page

Table A.6 (continued)

Feature	Source Column
gk_punches_p90	gk_punches

A.2.3 Statistical Baseline Dataset

Table A.7 lists the non-identifier features that compose the statistical baseline dataset.

Table A.7: Non-identifier features comprising the statistical baseline dataset

Feature	Feature
minutes_played	yellow_cards
red_cards	rating
dispossessed	hit_woodwork
penalties_won	offsides
penalties_conceded	backward_passes
turnovers	pass_accuracy
dribble_success_rate	tackle_win_rate
aerial_win_rate	ground_duel_win_rate
long_ball_accuracy	cross_accuracy
shot_accuracy	through_ball_accuracy
xg_per_shot	xg_conversion_rate
key_pass_rate	big_chance_miss_rate
xg_diff	goals_p90
assists_p90	passes_successful_p90
passes_in_final_third_p90	dribbles_successful_p90
tackles_won_p90	aerials_won_p90
duels_won_p90	long_balls_successful_p90
crosses_successful_p90	through_balls_successful_p90
shots_total_p90	xg_total_p90

Continued on next page

Table A.7 (continued)

Feature	Feature
xg_on_target_p90	key_passes_p90
chances_created_p90	big_chances_created_p90
touches_p90	fouls_won_p90
fouls_committed_p90	dribbles_suffered_p90
offsides_provoked_p90	ball_recoveries_p90
clearances_total_p90	interceptions_p90
steals_p90	blocks_defensive_p90
shots_blocked_p90	possession_lost_p90
gk_saves_total_p90	gk_saves_inside_box_p90
gk_high_claims_p90	gk_punches_p90
days_since_last_match	matchday
weighted_point_diff	weighted_pos_diff
rating_roll3	rating_roll5
rating_roll3_std	rating_roll5_std
strength_diff_roll3	strength_diff_roll5
is_home	next_rating

A.3 Feature Space Configurations by Branch

A.3.1 Dynamic Features

Table A.8 enumerates all dynamic (per-match) features fed into the LSTM branch. The core set of 46 features is shared across all three positional branches; the remaining 7 are position-specific, arising from role-specific statistics that are either structurally absent or uninformative for certain positions. A tick (✓) indicates that the feature is present in that branch.

Table A.8: Dynamic features by positional branch (MF, DF, FW)

Feature	**MF**	**DF**	**FW**
minutes_played	✓	✓	✓
yellow_cards	✓	✓	✓
rating	✓	✓	✓
dispossessed	✓	✓	✓
hit_woodwork	✓		✓
offsides			✓
penalties_won	✓	✓	✓
penalties_conceded	✓	✓	✓
backward_passes	✓	✓	✓
turnovers	✓	✓	✓
pass_accuracy	✓	✓	✓
dribble_success_rate	✓	✓	✓
tackle_win_rate	✓	✓	✓
aerial_win_rate	✓	✓	✓
ground_duel_win_rate	✓	✓	✓
long_ball_accuracy	✓	✓	✓
cross_accuracy	✓	✓	✓
shot_accuracy	✓	✓	✓
through_ball_accuracy	✓	✓	✓
xg_per_shot	✓	✓	✓
xg_conversion_rate	✓	✓	✓
key_pass_rate	✓	✓	✓
big_chance_miss_rate	✓	✓	✓
xg_diff	✓	✓	✓
goals_p90	✓	✓	✓
assists_p90	✓	✓	✓

Continued on next page

Table A.8 (continued)

Feature	MF	DF	FW
shots_total_p90	✓	✓	✓
xg_total_p90	✓	✓	✓
xg_on_target_p90	✓	✓	✓
key_passes_p90	✓	✓	✓
big_chances_created_p90	✓	✓	✓
chances_created_p90	✓		✓
touches_p90	✓	✓	✓
passes_successful_p90	✓	✓	✓
passes_in_final_third_p90	✓	✓	✓
long_balls_successful_p90	✓	✓	✓
crosses_successful_p90	✓	✓	✓
through_balls_successful_p90	✓		
dribbles_successful_p90	✓	✓	✓
tackles_won_p90	✓	✓	✓
aerials_won_p90	✓	✓	✓
duels_won_p90	✓	✓	✓
interceptions_p90	✓	✓	✓
steals_p90	✓	✓	✓
clearances_total_p90	✓	✓	✓
ball_recoveries_p90	✓	✓	✓
shots_blocked_p90	✓		✓
blocks_defensive_p90		✓	
offsides_provoked_p90		✓	
fouls_won_p90	✓	✓	✓
fouls_committed_p90	✓	✓	✓
dribbles_suffered_p90	✓	✓	✓
possession_lost_p90	✓	✓	✓

Continued on next page

Table A.8 (continued)

Feature	**MF**	**DF**	**FW**
Total	**50**	**48**	**50**

A.3.2 Static Features

Static features are fed into the MLP branch of the hybrid architecture. Unlike the dynamic features, the static feature set is *identical* across all positional branches. The distinction lies instead in the model *variant*. The Statistical Baseline configuration uses 5 match-level features, while the three context variants (Stats-Context, Article-Context, Full-Context) extend this with 14 LLM-derived contextual scores, totalling 19 static features.

Table A.9: Static features by model configuration

Feature	**Baseline**	**Context variants** (Stats / Article / Full)
`is_home`	✓	✓
`days_since_last_match`	✓	✓
`matchday`	✓	✓
`weighted_point_diff`	✓	✓
`weighted_pos_diff`	✓	✓
`fatigue`		✓
`sustainability`		✓
`role_stability`		✓
`confidence`		✓
`rate_of_change`		✓
`match_stakes`		✓
`match_difficulty`		✓
`fatigue_next`		✓
`sustainability_next`		✓
`role_stability_next`		✓

Continued on next page

Table A.9 (continued)

Feature	**Baseline**	**Context variants** (Stats / Article / Full)
`confidence_next`		✓
`rate_of_change_next`		✓
`match_stakes_next`		✓
`match_difficulty_next`		✓
Total	**5**	**19**

Appendix B

Prompt Schema References

B.1 Player History Column Schema

When constructing the `Stats-Context` and `Full-Context` prompts, the statistical history for each player was provided to the LLM as a comma-separated string. The first row of this block is the header schema, which maps directly to the features engineered in Section 4.1.3. The columns included in the input vector are listed below:

Table B.1: Exact column sequence for the Stats-Context string

Feature	**Feature**
`date`	`match_id`
`minutes`	`yellow_cards`
`red_cards`	`rating`
`dispossessed`	`hit_woodwork`
`penalties_won`	`offsides`
`penalties_conceded`	`backward_passes`
`turnovers`	`pass_accuracy`
`dribble_success_rate`	`tackle_win_rate`
`aerial_win_rate`	`ground_duel_win_rate`
`long_ball_accuracy`	`cross_accuracy`
`shot_accuracy`	`through_ball_accuracy`
`xg_per_shot`	`xg_conversion_rate`

Continued on next page

Table B.1 (continued)

Feature	Feature
key_pass_rate	big_chance_miss_rate
xg_diff	goals_p90
assists_p90	passes_successful_p90
passes_in_final_third_p90	dribbles_successful_p90
tackles_won_p90	aerials_won_p90
duels_won_p90	long_balls_successful_p90
crosses_successful_p90	through_balls_successful_p90
shots_total_p90	xg_total_p90
xg_on_target_p90	key_passes_p90
chances_created_p90	big_chances_created_p90
touches_p90	fouls_won_p90
fouls_committed_p90	dribbles_suffered_p90
offsides_provoked_p90	ball_recoveries_p90
clearances_total_p90	interceptions_p90
steals_p90	blocks_defensive_p90
shots_blocked_p90	possession_lost_p90
gk_saves_total_p90	gk_saves_inside_box_p90
gk_high_claims_p90	gk_punches_p90
days_since_last_match	weighted_point_diff
weighted_pos_diff	rating_roll3
rating_roll5	rating_roll3_std
rating_roll5_std	strength_diff_roll3
strength_diff_roll5	

B.2 Baseline Prompt Template

The template below was used by the Stats-Context variant to score all seven constructs from scratch. Dynamic fields are shown in angle brackets; the `<match context>` block corresponds

to the key-value listing illustrated in Listing 4.1, and the `<player history>` block to the statistical history format illustrated in Listing 4.2.

```
You are a football performance analyst. Score pre-match signals
in [0,1] using ONLY the data below. Use 0.5 when evidence is
weak. Do not invent facts.

### PLAYER METRICS (from match history)
fatigue: physical load entering this match (minutes, action
volume, fixture congestion, rest gap). 1=heavily loaded or
depleted, 0=fully rested and fresh.

sustainability: reliability of recent output level. 1=performance
is stable and repeatable across recent matches, 0=current average
is inflated by a single outlier in an otherwise poor or erratic
run.

role_stability: consistency of deployment. 1=playing usual
position with habitual minutes, 0=shifted to a new role or
playing atypical minutes relative to their season pattern.

confidence: current psychological state and sustained performance
level. 1=player is in a positive momentum state (goal
involvement, coach trust, things going their way across recent
matches), 0=player is struggling (dropped, criticised, visibly
poor run, low morale).

rate_of_change: velocity and recency of form movement,
independent of current level. 1=large sudden positive shift in
the most recent match(es) relative to prior trend, 0=sharp sudden
negative shift. Use 0.5 when form is flat or change is gradual.
If a player block says "No previous matches available", output
0.5 for all metrics.

### TEAM METRICS (from match context)
match_stakes: how much this result matters (title race,
relegation, cups). 1=critical, 0=dead rubber.
match_difficulty: opponent strength relative to THIS team.
1=THIS team is heavy underdog, 0=THIS team is heavy favourite,
0.5=evenly matched.

### OUTPUT FORMAT
Raw JSON only. No markdown.
{"team":{"match_stakes":float|null,"match_difficulty":float|null,
"reasoning":"str"},"players":[{"player_id":"str",
"player_name":"str","fatigue":float|null,
"sustainability":float|null,"role_stability":float|null,
"confidence":float|null,"rate_of_change":float|null,
"reasoning":"str"}]}

### MATCH CONTEXT
<match metadata: key-value pairs>

### PLAYER HISTORY
<per-player statistical history blocks>
```

Listing B.1: Baseline prompt template (Stats-Context)

B.3 Adjustment Prompt Template

The template below was used by the Article-Context and Full-Context variants to update the Stats-Context prior. The two variants share the same structural skeleton. Blocks marked `[Full-Context only]` appear exclusively when `include_match_context` and `include_player_history` are enabled. The `TASK` and `RULES` sections also carry variant-specific text, as indicated by the inline labels. The listing shows the Full-Context version in full and the Article-Context version omits the `PLAYER HISTORY` evidence blocks and uses the simpler `TASK` and `RULES` variants.

```
You are a football performance analyst updating pre-computed
signals.

### TASK
[Article-Context]
Baseline signals were computed from historical stats.
Adjust ONLY when articles provide clear, specific evidence. Do
not recompute from scratch. Return null for any metric you are
not changing.

[Full-Context]
Baseline signals were computed from historical stats. Additional
context provided: articles, player history.
Adjust using ALL sources. Cross-reference them: article signals
carry more weight when corroborated by history or context.
Return null for any metric where no source justifies a change.

### SIGNAL REFERENCE
fatigue: raise on reports of fixture congestion, travel burden,
or injury scare; lower on confirmed rest or rotation.

sustainability: adjust only if articles reveal the recent output
level is misleading -- e.g., a standout performance against weak
opposition, or a return from injury inflating a short sample.

role_stability: adjust on confirmed positional change, unexpected
absence, or atypical minutes. Weight stronger if player history
already shows irregular deployment.

confidence: adjust on clear reported momentum signals -- goal
involvement, coach praise or public criticism, dropped from
starting XI, visible morale indicators. Weight stronger if player
history already shows a sustained trend in the same direction.

rate_of_change: adjust only for sharp, sudden shifts -- injury
return ramping up abruptly, a standout match after a long poor
```

```
run, or a public form collapse. Do not adjust for gradual trends;
those belong to confidence.
match_stakes: adjust if article reveals new context (must-win
confirmed, points gap changed, cup scenario).
match_difficulty: opponent strength relative to THIS team.
Rarely changes -- only on major opponent absences.

### MATCH
- Match ID: <match_id> | Team: <team_name>
- Opponent: <opponent_name> | Date: <match_date>
- Venue: <home|away>

### SQUAD
<bulleted list: [player_id] Player Name (position)>

### BASELINE TEAM SIGNALS
<prior match_stakes and match_difficulty scores>

### BASELINE PLAYER SIGNALS
<prior per-player scores for all five constructs>

### MATCH CONTEXT
<match metadata key-value listing>

### PLAYER HISTORY  [Full-Context only]
<per-player statistical history blocks>

### RULES
[Article-Context]
- Values in [0,1] only. Null = no change.
- Use exact player_id and player_name from the squad list.
  Ignore players not in squad.
- Reasoning must explain each changed metric with previous value,
  new value, and concrete article evidence.
- For each changed metric:
    <metric>: <old> -> <new> | evidence: <fact/quote>
    | source: articles | why: <justification>
- If no metric changed:
    no_change: insufficient evidence to override baseline.

[Full-Context]
- Cross-reference all sources; corroborated signals are stronger.
- Values in [0,1] only. Null = no change.
- Use exact player_id and player_name from the squad list.
  Ignore players not in squad.
- Reasoning must explain each changed metric with previous value,
  new value, and concrete evidence.
- For each changed metric:
    <metric>: <old> -> <new> | evidence: <fact/quote>
    | source: <articles|match_context|player_history|combined>
    | why: <justification>
- If no metric changed:
    no_change: insufficient evidence to override baseline.

### OUTPUT FORMAT
Raw JSON only. No markdown.
{"team":{"match_stakes":float|null,"match_difficulty":float|null,
```

```
"reasoning":"str"},"players":[{"player_id":"str",
"player_name":"str","fatigue":float|null,
"sustainability":float|null,"role_stability":float|null,
"confidence":float|null,"rate_of_change":float|null,
"reasoning":"str"}]}

### ARTICLES
<phased article window>
```

Listing B.2: Adjustment prompt template (Article-Context and Full-Context)

Appendix C

Reliability Audit Sample

This appendix lists the thirty highest-deviation cases that formed the reliability audit sample defined in Section 4.2.5. Each row is a single feature adjustment, selected among the thirty largest absolute deviations ($|\Delta|$) the Article-Context variant produced relative to its Stats-Context prior across the five player-level contextual features. The Baseline column gives the Stats-Context prior value, the Article column the value after the article-driven adjustment, and the $|\Delta|$ column their absolute difference. Because the selection is made over individual feature adjustments rather than whole player-matches, a single player may appear in more than one row.

Table C.1: The 30 player-match cases with the largest absolute deviation ($|\Delta|$) selected for the reliability audit

Player	**Pos**	**Feature**	**Baseline**	**Article**	$\lvert\Delta\rvert$
Sikou Niakaté	DF	Role stability	0.90	0.10	0.80
André Lacximicant	FW	Confidence	0.30	0.85	0.55
Nigel Thomas	FW	Fatigue	0.40	0.90	0.50
André Lacximicant	FW	Rate of change	0.30	0.80	0.50
Yanis Begraoui	FW	Confidence	0.35	0.85	0.50
Vasco Lopes	FW	Fatigue	0.50	0.00	0.50
Vasco Lopes	FW	Rate of change	0.65	0.20	0.45
Sikou Niakaté	DF	Fatigue	0.65	0.20	0.45
Rodrigo Mora	FW	Fatigue	0.65	0.20	0.45

Continued on next page

Table C.1 (continued)

Player	**Pos**	**Feature**	**Baseline**	**Article**	$\lvert\Delta\rvert$
Yanis Begraoui	FW	Confidence	0.45	0.90	0.45
Vangelis Pavlidis	FW	Confidence	0.40	0.85	0.45
Henrique Pereira	FW	Fatigue	0.25	0.70	0.45
Jalen Blesa	FW	Confidence	0.50	0.90	0.40
Francisco Trincão	FW	Confidence	0.35	0.75	0.40
Beni Mukendi	MF	Role stability	0.80	0.40	0.40
Vasco Lopes	FW	Role stability	0.50	0.10	0.40
Vasco Lopes	FW	Rate of change	0.50	0.10	0.40
Vasco Lopes	FW	Fatigue	0.40	0.80	0.40
Gil Dias	FW	Fatigue	0.40	0.80	0.40
Vítor Gomes	MF	Role stability	0.30	0.70	0.40
Vinícius Zanocelo	MF	Fatigue	0.70	0.30	0.40
Xeka	MF	Role stability	0.30	0.70	0.40
José Gomes	MF	Role stability	0.70	0.30	0.40
Ricardo Horta	FW	Confidence	0.45	0.85	0.40
Miguel Reisinho	MF	Fatigue	0.70	0.30	0.40
Enzo Barrenechea	MF	Fatigue	0.70	0.30	0.40
Gustavo Silva	FW	Fatigue	0.30	0.70	0.40
Martim Fernandes	DF	Fatigue	0.55	0.20	0.35
Vasco Lopes	FW	Role stability	0.55	0.20	0.35
Yanis Begraoui	FW	Confidence	0.30	0.65	0.35

Appendix D

Predictive Results

This appendix collects the full per-bin error breakdowns and the predicted-versus-actual scatter plots for all fifteen position and variant combinations evaluated in Chapter 4. Section D.1 reports the per-bin RMSE for every configuration, the exact values summarised by the curves in Figure 4.6. The remaining sections collect the scatter plots: each shows the held-out test predictions against the observed ratings, together with the identity line and a band of ± 0.5 around it. Across every configuration the predictions concentrate into a narrow central band while the observed ratings span the full $[6.0, 8.0]$ range, the regression-toward-the-mean behaviour discussed in Section 4.5.2. The scatter plots are grouped by model configuration, in the order the configurations are introduced in Table 4.10, and ordered by position within each group.

D.1 Per-Bin RMSE Breakdowns

The following tables give the per-bin RMSE for every model configuration, computed over rating bins of width 0.1 across the $[6.0, 8.0]$ target range. Bold marks the lowest RMSE in each bin, and n is the number of test rows in the bin. These are the exact values summarised by Figure 4.6 in Section 4.5.2.

Table D.1: Per-bin RMSE for forwards across the [6.0, 8.0] rating range. Bold marks the lowest RMSE in each bin and n is the number of test rows in the bin.

Bin	n	Average	Baseline	Stats-Context	Article-Context	Full-Context
6.0–6.1	13	**0.6605**	0.8984	0.8956	0.8739	0.9217
6.1–6.2	13	**0.6364**	0.8384	0.8721	0.8440	0.9314
6.2–6.3	38	**0.4711**	0.6801	0.6440	0.6339	0.6192
6.3–6.4	53	**0.3841**	0.6096	0.5857	0.5764	0.5624
6.4–6.5	119	**0.2633**	0.3981	0.4115	0.3951	0.3837
6.5–6.6	133	**0.2779**	0.4136	0.4104	0.3936	0.4097

Continued on next page

Table D.1 (continued)

Bin	*n*	Average	Baseline	Stats-Context	Article-Context	Full-Context
6.6–6.7	114	**0.2046**	0.3512	0.3206	0.3032	0.3333
6.7–6.8	73	**0.2194**	0.3083	0.2784	0.2562	0.3398
6.8–6.9	70	**0.2648**	0.3152	0.2952	0.2757	0.3429
6.9–7.0	35	**0.2552**	0.2995	0.2623	0.2694	0.3491
7.0–7.1	48	0.3846	0.3379	**0.2655**	0.2698	0.3768
7.1–7.2	31	0.3950	0.4131	**0.3240**	0.3395	0.4029
7.2–7.3	25	0.4673	0.4210	**0.3637**	0.3690	0.4259
7.3–7.4	21	0.5768	0.5009	**0.4321**	0.4530	0.5443
7.4–7.5	19	0.6251	0.6070	**0.5135**	0.5371	0.6270
7.5–7.6	10	0.7251	0.5559	**0.5479**	0.5975	0.6783
7.6–7.7	9	0.7944	0.7890	**0.7039**	0.7363	0.7557
7.7–7.8	9	0.9997	0.8816	**0.8123**	0.8129	0.9414
7.8–7.9	5	0.9808	0.8010	**0.6621**	0.7092	0.6950
7.9–8.0	7	1.1214	0.7989	0.7993	0.8264	**0.7915**

Table D.2: Per-bin RMSE for defenders across the [6.0, 8.0] rating range. Bold marks the lowest RMSE in each bin and *n* is the number of test rows in the bin.

Bin	*n*	Average	Baseline	Stats-Context	Article-Context	Full-Context
6.0–6.1	18	0.7527	0.8196	0.4765	**0.4354**	0.4657
6.1–6.2	38	0.6302	0.7260	0.5376	0.5334	**0.5153**
6.2–6.3	40	0.5776	0.7425	0.4812	0.4840	**0.4659**
6.3–6.4	62	0.4857	0.6621	0.3952	0.3807	**0.3691**
6.4–6.5	80	0.3404	0.4780	0.2787	0.2731	**0.2528**
6.5–6.6	87	**0.2627**	0.4430	0.2771	0.2920	0.2755
6.6–6.7	97	**0.2165**	0.3765	0.2492	0.2611	0.2441
6.7–6.8	79	**0.2701**	0.3990	0.2785	0.2932	0.2773
6.8–6.9	74	**0.2195**	0.3562	0.2744	0.2941	0.2778
6.9–7.0	71	**0.2681**	0.3532	0.3480	0.3628	0.3613
7.0–7.1	73	**0.3257**	0.3688	0.4269	0.4459	0.4338
7.1–7.2	51	**0.3433**	0.4056	0.4682	0.4714	0.4763
7.2–7.3	29	0.4998	**0.4559**	0.5551	0.5681	0.5661
7.3–7.4	26	0.5790	**0.5409**	0.6437	0.6364	0.6489
7.4–7.5	40	0.5501	**0.4613**	0.5966	0.6059	0.6276
7.5–7.6	29	0.7988	**0.6106**	0.8230	0.8408	0.8348
7.6–7.7	17	0.7705	**0.6750**	0.8628	0.8968	0.8914
7.7–7.8	8	0.9911	**0.7356**	0.9975	0.9947	1.0145

Continued on next page

Table D.2 (continued)

Bin	*n*	Average	Baseline	Stats-Context	Article-Context	Full-Context
7.8–7.9	5	0.8505	**0.7863**	0.9719	0.9385	0.9800
7.9–8.0	6	0.9135	**0.8757**	1.0700	1.0676	1.0988

Table D.3: Per-bin RMSE for midfielders across the [6.0, 8.0] rating range. Bold marks the lowest RMSE in each bin and *n* is the number of test rows in the bin.

Bin	*n*	Average	Baseline	Stats-Context	Article-Context	Full-Context
6.0–6.1	5	0.9071	0.9112	0.7960	**0.6545**	0.7425
6.1–6.2	26	0.6189	0.5900	**0.5110**	0.6174	0.5685
6.2–6.3	29	0.5799	0.5357	0.4881	0.5648	**0.4770**
6.3–6.4	71	**0.4483**	0.4870	0.4946	0.5780	0.4808
6.4–6.5	89	**0.3525**	0.3869	0.3853	0.4444	0.3646
6.5–6.6	131	0.3283	0.3315	0.3344	0.4113	**0.3063**
6.6–6.7	119	0.2597	0.2681	0.2504	0.3722	**0.2134**
6.7–6.8	90	0.2380	0.2361	0.2518	0.2662	**0.2205**
6.8–6.9	84	0.2165	0.2126	0.2738	0.2960	**0.2110**
6.9–7.0	64	0.2501	**0.2486**	0.3010	0.3150	0.2581
7.0–7.1	74	**0.2454**	0.2484	0.2984	0.3128	0.2564
7.1–7.2	43	0.3683	0.3602	0.3815	0.3950	**0.3549**
7.2–7.3	44	0.4347	0.3859	0.4110	0.4537	**0.3783**
7.3–7.4	32	0.5140	0.5226	0.5131	**0.4909**	0.5310
7.4–7.5	24	0.5441	**0.5377**	0.5654	0.5418	0.5434
7.5–7.6	22	0.7164	0.6183	0.6276	**0.6132**	0.6450
7.6–7.7	19	0.7552	0.7158	0.6875	**0.6605**	0.7100
7.7–7.8	7	0.8320	**0.7984**	0.8180	0.8155	0.8447
7.8–7.9	7	0.8483	0.8443	0.8272	0.8963	**0.8230**
7.9–8.0	6	**1.1349**	1.1752	1.1778	1.1656	1.1852

D.2 Average Baseline

Figure D.1: Predicted vs. actual performance scores for Defenders – Average Baseline.

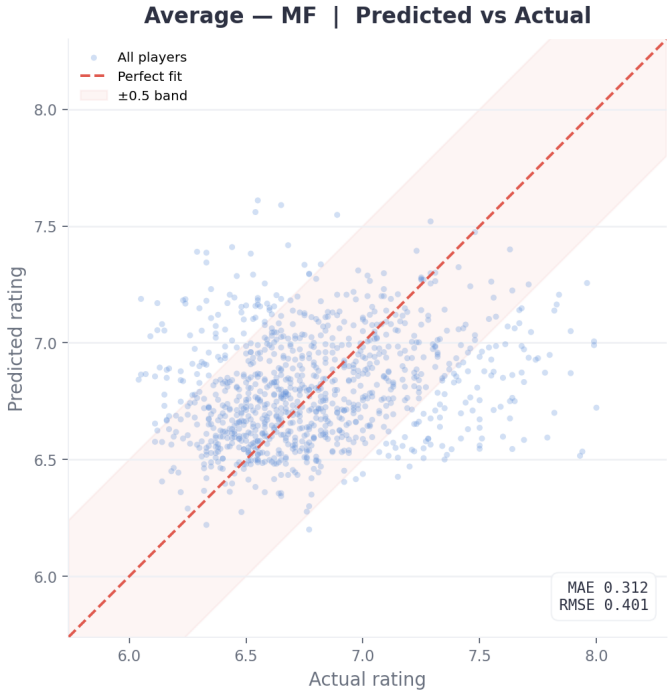

Figure D.2: Predicted vs. actual performance scores for Midfielders – Average Baseline.

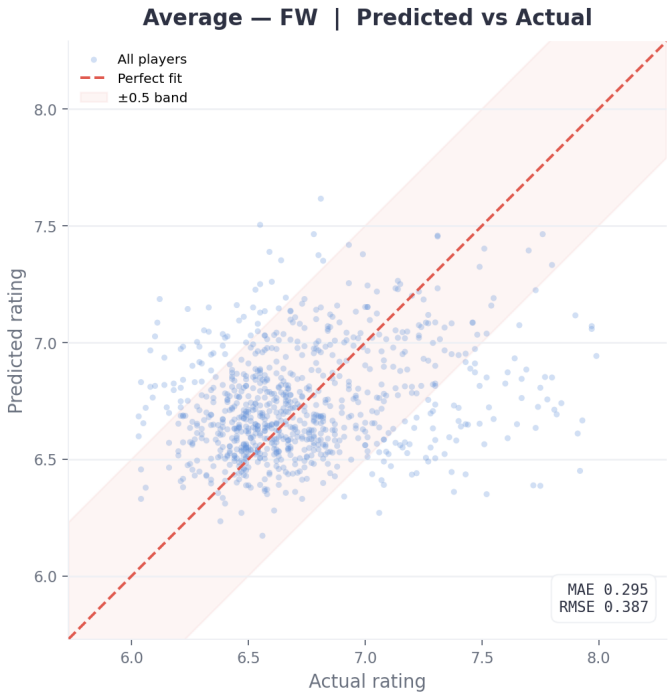

Figure D.3: Predicted vs. actual performance scores for Forwards – Average Baseline.

D.3 Statistical Baseline

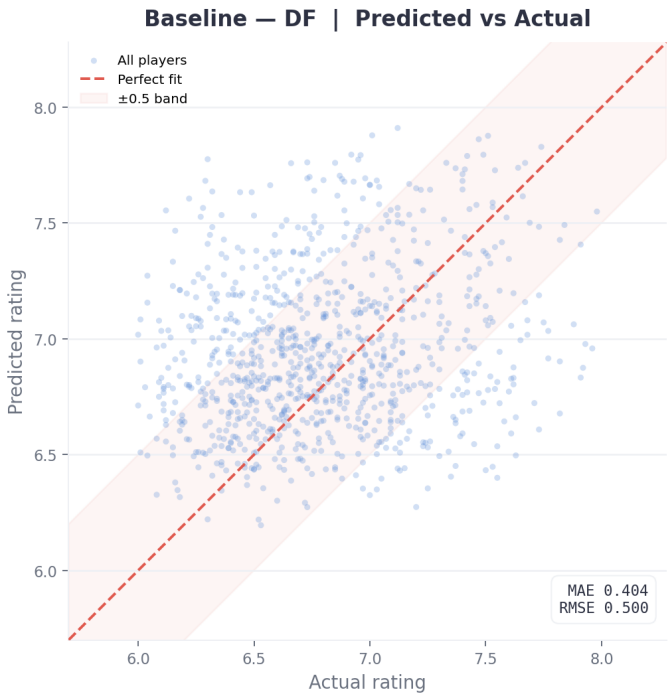

Figure D.4: Predicted vs. actual performance scores for Defenders – Statistical Baseline.

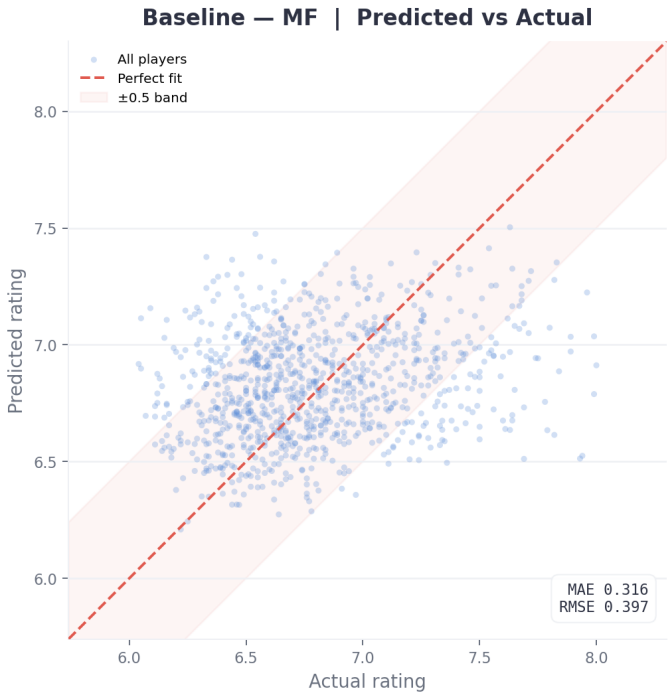

Figure D.5: Predicted vs. actual performance scores for Midfielders – Statistical Baseline.

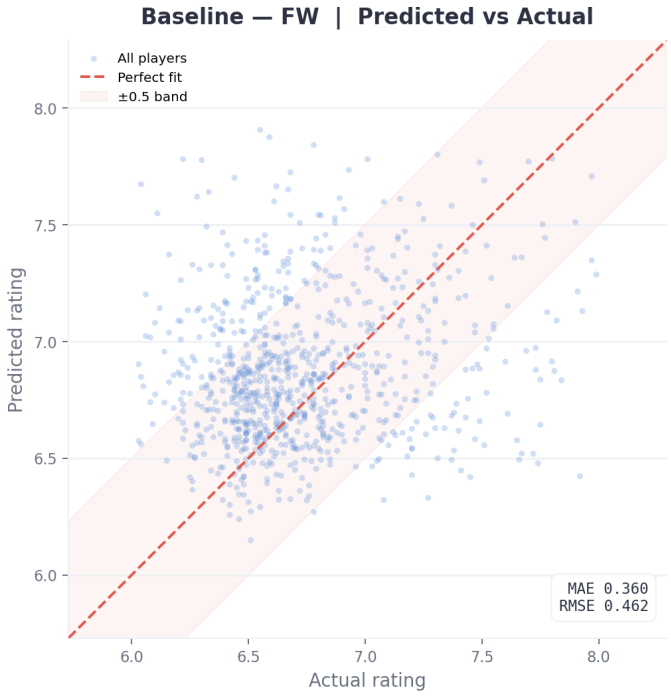

Figure D.6: Predicted vs. actual performance scores for Forwards – Statistical Baseline.

D.4 Stats-Context

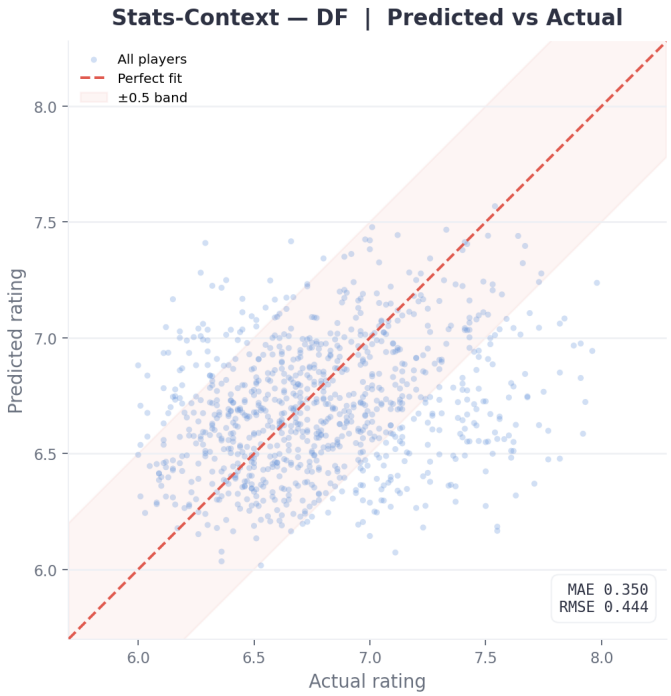

Figure D.7: Predicted vs. actual performance scores for Defenders – Stats-Context variant.

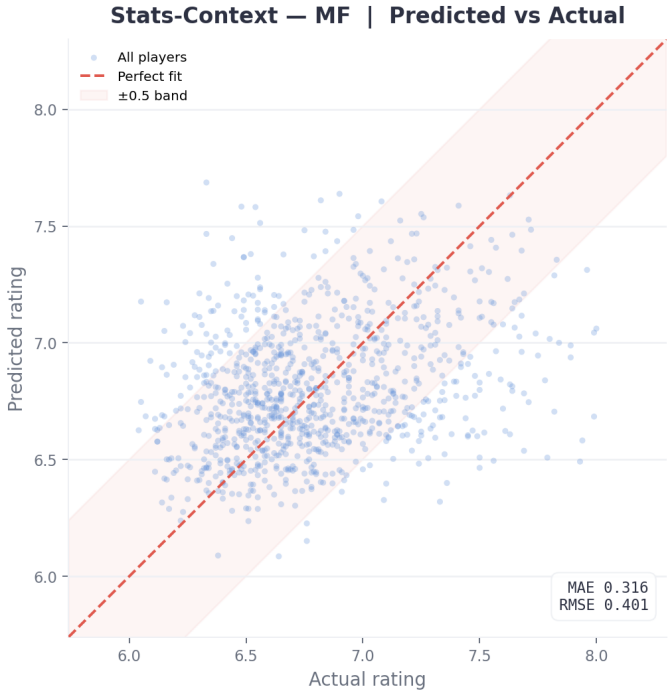

Figure D.8: Predicted vs. actual performance scores for Midfielders – Stats-Context variant.

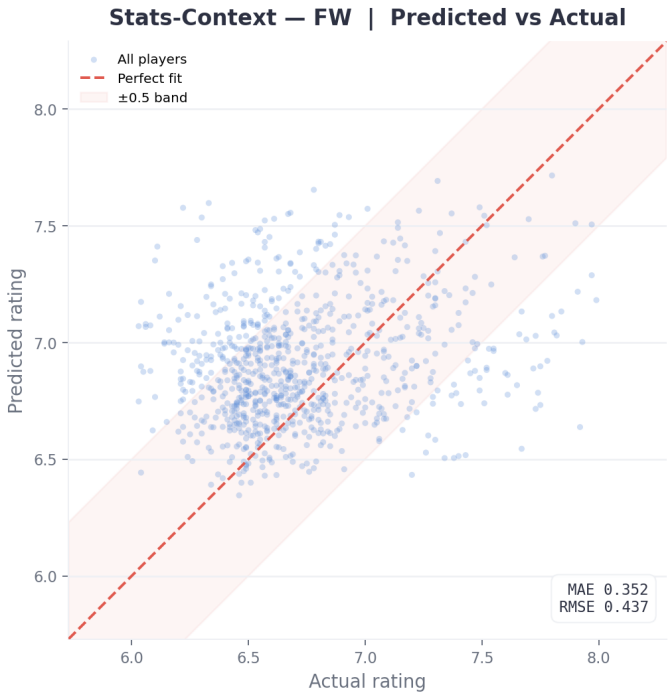

Figure D.9: Predicted vs. actual performance scores for Forwards – Stats-Context variant.

D.5 Article-Context

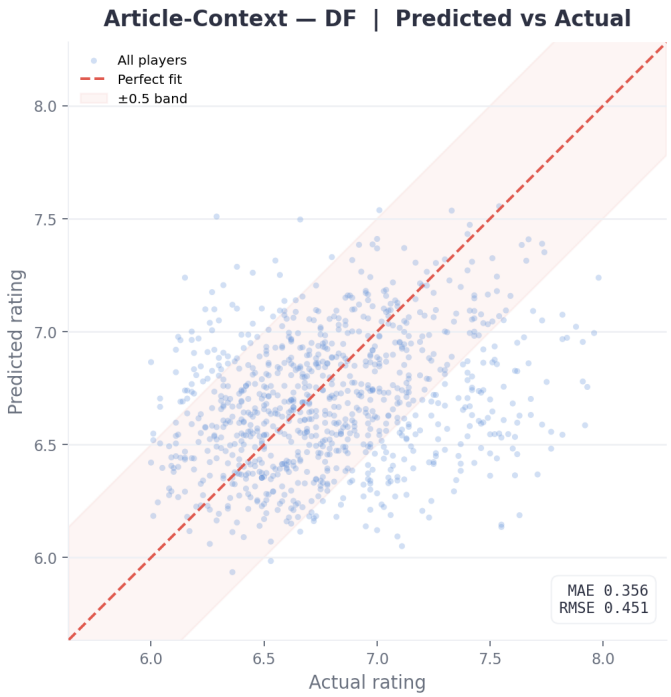

Figure D.10: Predicted vs. actual performance scores for Defenders – Article-Context variant.

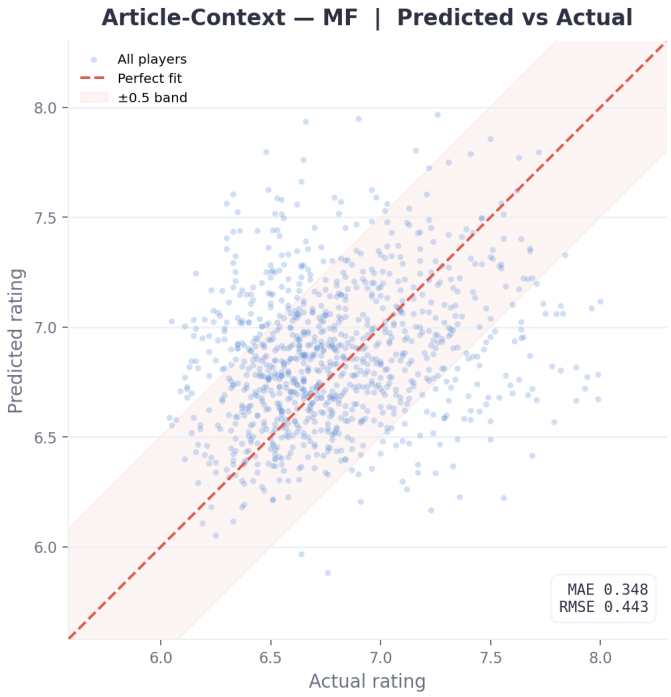

Figure D.11: Predicted vs. actual performance scores for Midfielders – Article-Context variant.

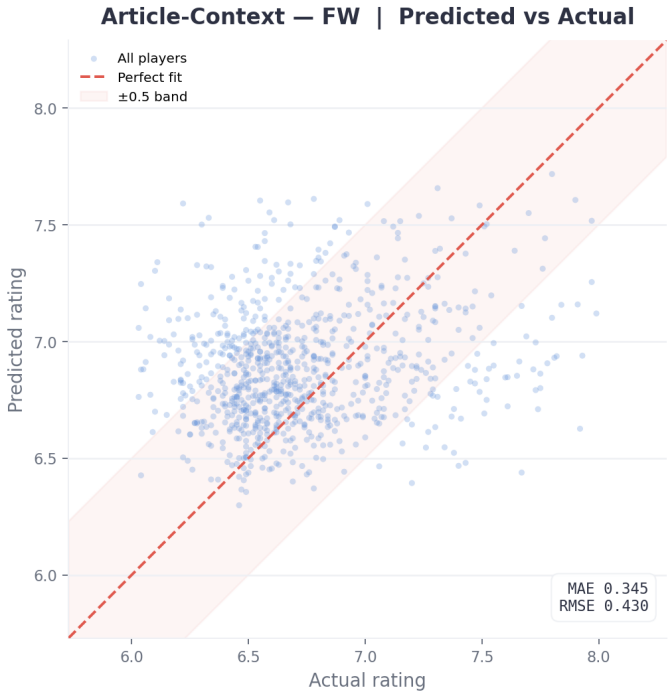

Figure D.12: Predicted vs. actual performance scores for Forwards – Article-Context variant.

D.6 Full-Context

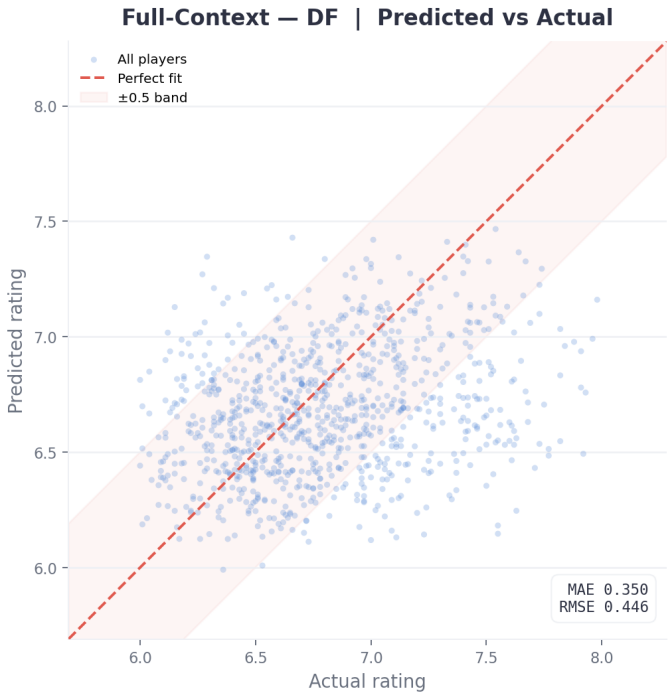

Figure D.13: Predicted vs. actual performance scores for Defenders – Full-Context variant.

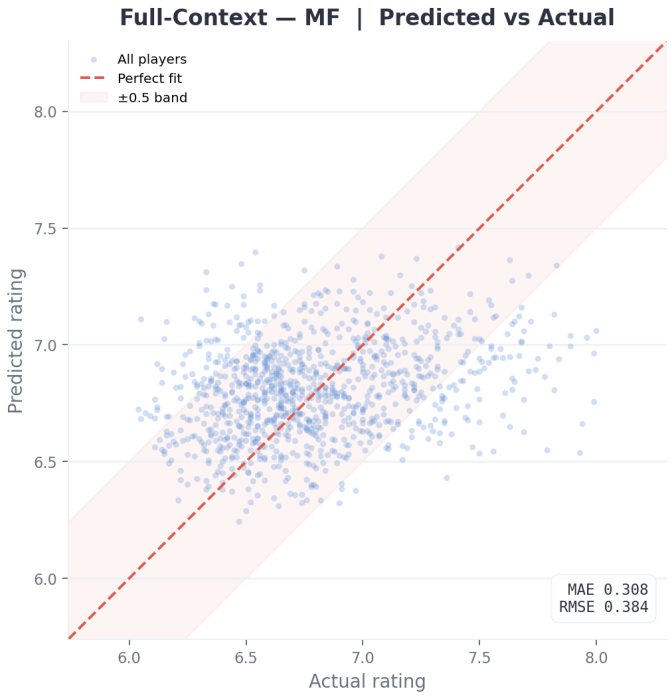

Figure D.14: Predicted vs. actual performance scores for Midfielders – Full-Context variant.

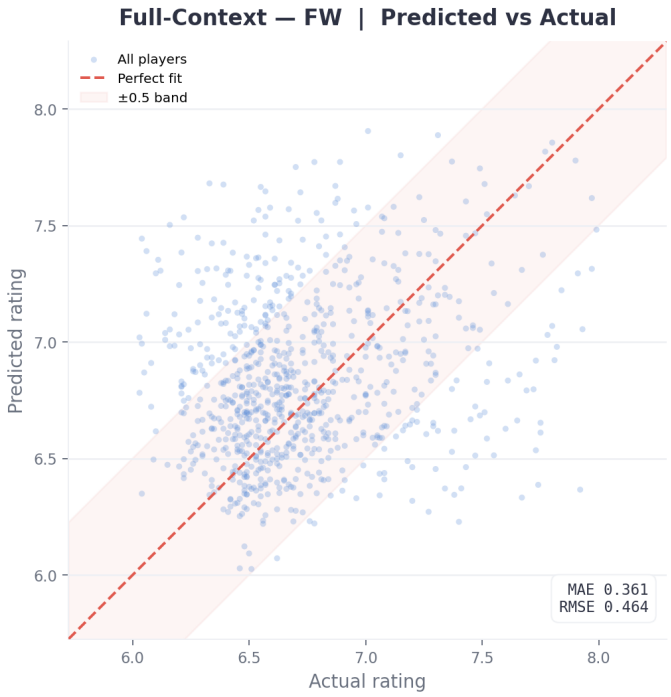

Figure D.15: Predicted vs. actual performance scores for Forwards – Full-Context variant.