POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

# A Critical Analysis of the Expected Goals (xG) Metric in Professional Soccer

Tesi di Laurea Magistrale in
Computer Science and Engineering - Ingegneria Informatica

**Luca Romano, 247714**

**Advisor:**
Prof. Fabrizio Pittorino

**Co-advisors:**
Alessandro Falcetta
Alessio Rossi
Carlo Simonelli

**Academic year:**
2024-2025

**Abstract:** Expected Goals (xG) models have become a central tool in modern football analytics, yet their exclusive focus on shot quality imposes structural limitations when used to evaluate team performance or predict match outcomes. This study conducts a three-stage analysis aimed at critically reassessing the role and reliability of xG using publicly available data.
A first highly effective xG model was developed by expanding the engineered feature set, leveraging freeze-frame positional context and contextual descriptors, to achieve top-tier accuracy using publicly available data. Once it was ensured that the model performed better than current models, the operational framework was made public to facilitate the model's interpretability and the model's limitations were then analyzed and discussed. To go beyond these limitations, a complementary event-based performance metric was then introduced, aggregating attacking and defensive behaviours over the full match and capturing dimensions of performance that go beyond shot-based information. Finally, a combined approach was evaluated by integrating team-level xG features into the event-based model.
Using historical data from men's elite competitions, the event-based model was shown to slightly outperform xG in predicting match outcomes, while the combined model achieves the highest overall effectiveness. These results indicate that xG alone cannot fully explain match dynamics, but remains highly informative when embedded within broader representations of team behaviour. The findings highlight the need for more comprehensive performance metrics and clarify the complementary roles of shot-based and event-based information in football analytics.

**Key-words:** Expected Goals (xG), Football analytics, Event data, Spatial features

## 1. Introduction

In recent years, football performance has dramatically changed with the growing presence of event-based and positional data [11]. This transformation pushed performance analysis towards quantitative data-driven methods that are increasingly used by professional football clubs, media and independents analysts [12]. Among the many metrics developed, the *Expected Goals* (xG) model has emerged as a key concept to evaluate shot quality and team performance [15, 19].

Expected Goals (xG) is a probabilistic metric that estimates the likelihood that a given shot will result in a goal. For each shot, an xG model outputs the probability (expressed in a range from 0 to 1) based on a set of features that summarises the context in which the shot is taken. Generally, the probability represents the percentage of times the same type of shot (with the same contextual variables) has been scored historically (in the available data) [19].
The objectives of this study will be presented as a list of Research Questions (RQ).

## 1.1. Research Questions

This study addresses the following research questions:

- **RQ1:** How can xG models be improved using publicly available data?
- **RQ2:** Can a public and accessible operational framework be defined to standardise xG construction and evaluation?
- **RQ3:** What are xG's intrinsic statistical and conceptual limitations, and how is it commonly misused or misinterpreted in practice?
- **RQ4:** How can xG be complemented with additional data-driven indicators to obtain a more complete representation of performance?

A brand new xG model (RQ1) was built by integrating freeze-frame positional information and engineered spatial context, all derived from public StatsBomb data [21], together with contextual descriptors and domain-specific football knowledge, with the aim of achieving a level of accuracy in line with or superior to those currently present in the state-of-the-art. In doing so, the study delivers a carefully engineered xG estimator that enables a more reliable and accurate use of the metric in practical settings, supporting practitioners, analysts, and media commentators who routinely base evaluations on xG.
A further motivation for this study arises from the lack of an operationally consistent framework [18] for defining, modelling, and interpreting xG. Despite its widespread use, xG is computed through heterogeneous pipelines that differ in feature sets, preprocessing rules, modelling choices, and evaluation criteria, leading to substantial variability in the resulting values and in their practical interpretation [23]. Establishing clearer methodological standards is essential to improve accuracy, ensure comparability across studies, and support consistent usage among analysts and practitioners. In this regard (RQ2), the modelling pipeline developed makes each stage of the xG construction process (data preparation, feature engineering, spatial context extraction, cross-validation design, and model evaluation) explicit and public, providing a transparent and reproducible reference framework that can help clarify how xG values emerge from specific modelling choices.
The xG metric has been adopted in many football contexts, but its practical meaning and reliability remain contested: xG is often interpreted as a deterministic indicator of performance rather than a probabilistic estimate [2, 19], leading to misinterpretations and, in practice, to an overreliance on the metric in applied analysis and media commentary [5]. For example, simply summing shot-level xG values and treating the result as a definitive match-level measure can be misleading: ten shots with an xG value of 0.099 are not the same of a single shot with an xG value of 0.99, both statistically and conceptually.
While much research has focused on improving predictive accuracy of xG metric through richer datasets or more sophisticated machine-learning pipelines [2, 13, 15], less attention has been devoted to understanding its structural limitations [17]. In particular, using shot-level xG estimates to predict match-level outcome estimates remains inherently noisy: aggregating chance probabilities to infer an expected winner provides only moderate reliability, especially in tightly balanced games [6, 22], as they often fails to consider substantial game context unrelated to shots. In fact, their predictive effectiveness improves when additional match information and temporal structure are incorporated [8].
The effectiveness of xG models and the investigation of the main factors explains why xG can fail to provide reliable informative value in practice (RQ3).
Finally, a complementary event-based model and a combined formulation that integrates team-level xG information with match-level event signals (RQ4) were introduced to enrich performance analysis at match-level, providing a reliable performance metric for practitioners and analysts who need a model capable of completing the information necessary to carry out analyses at that level, which xG metric alone was proven not to provide sufficiently.
This work also incorporates insights from professional sports data scientists, ensuring that the proposed methodology remains grounded in practical decision-making contexts in elite football.

## 1.2. Models' introduction

This study was developed through a three-stage analysis that corresponded to the creation of three new models. First, we develop an enhanced xG model (xG-SpatialAware model) based on engineered features derived from

freeze-frame positional information, contextual descriptors, and domain-specific football knowledge, with the goal of improving predictive effectiveness using publicly available data. Beyond improving accuracy, this model also serves as a rigorous reference point to analyse systematic behaviours and residual errors that persist even under a rich, spatially informed feature set.
Second, motivated by these findings, a complementary event-based performance model (Event-based model) was introduced to capture attacking and defensive behaviours that xG systematically overlooks. This model aggregates match events over the full ninety minutes and was evaluated on the basis of its ability to predict match outcomes.
Finally, the two approaches were combined by incorporating team-level xG features into the event-based model, allowing the degree to which the two perspectives encode complementary information about match performance to be quantified. This integrated approach (Combined model) yields the highest predictive accuracy, demonstrating that shot-based and event-based information represent distinct but complementary dimensions of team performance.

# 2. Related literature

This section aims to summarize the existing xG literature to date. First, we examine early definitions and the conceptual roots of its intended objectives, and discuss how the metric has evolved from its initial statistical formulations to the modern shot-based models adopted in contemporary analysis. We then focus on the advances made to improve xG estimation, with a particular focus on approaches that incorporate positional or still-image information to better represent the shooting context. The second part of the chapter moves to critical perspectives, summarizing works that highlight the inherent limitations of xG (both statistical and conceptual) and examining proposed remedies, including more comprehensive modeling assumptions and complementary information sources. Finally, we examine alternative metrics that build on the same probabilistic intuition as xG, while addressing broader notions of performance that go beyond shot quality.

## 2.1. Foundations: definition and historical evolution of xG

In the literature, expected goals (xG) are defined as a probabilistic model of shot success: an estimate of the probability that a given attempt will result in a goal, based on contextual variables describing the shot. Early academic contributions, building on this concept, used positional features to study shot efficiency and expected score [19]. A representative milestone in this direction is the work by Lucey et al., which models shot outcomes using strategic features extracted from tracking sequences and highlights the role of contextual information beyond shot location alone [13]. In parallel, Eggels et al. proposed a probabilistic approach that estimates scoring probabilities for individual opportunities and uses them to infer expected match outcomes, providing an early example of connecting shot-level probabilities to match-level reasoning [6].

## 2.2. xG improvements: richer datasets, machine learning techniques and probabilistic corrections

Recent research has shown that xG effectiveness can be improved by enriching shot descriptions with positional context (e.g. the location of defenders and the goalkeeper at the moment of the attempt). Compared to location-only models, synchronized positional data enable features that more directly represent defensive pressure, occlusion, and the effective shooting window, thereby providing a more faithful approximation of shot difficulty [2, 15]. This line of work has been further improved by efforts focusing on interpretability [4] and on probabilistic corrections for player and positional effects [20].
Taken together, these improvements have led to consistent gains in xG accuracy, suggesting that modern, context-rich formulations are approaching the upper bound attainable from shot-time information.

## 2.3. xG limitations

Despite its widespread adoption, the literature consistently emphasises that xG should be interpreted as a probabilistic estimate of shot success rather than as a deterministic statement about performance. This distinction is often ignored in practice, where xG values are used to support definitive narratives about which team deserved to win or about a player's finishing ability, even though short-horizon deviations from expectation are fully compatible with statistical variance [3]. As a consequence, several studies caution that xG-based conclusions are highly sensitive to sample size: at match level, the number of shots is typically small, which makes both aggregated xG totals and derived inferences inherently noisy [22].

A closely related limitation concerns the aggregation of shot-level probabilities to draw match-level conclusions. While summing xG across attempts yields an expected number of goals, this quantity does not directly translate into an outcome probability and can support misleading interpretations when treated as a stand-alone indicator of match dominance [3, 6]. More fundamentally, xG is defined on observed shots and therefore cannot represent dimensions of performance that affect outcomes without necessarily manifesting as shot quality, such as territorial control, sustained pressure, or the ability to generate and suppress shot volume over long phases of play. Recent work shows that match outcome prediction can benefit from incorporating additional match information and temporal structure beyond shot quality alone, highlighting the incomplete nature of an xG-only representation [8].

The literature also reports conceptual and methodological issues arising when xG is used to evaluate players' finishing ability. In particular, biases and modelling choices can confound attempts to separate finishing ability from shot quality, since the estimated shot probability is itself affected by the available covariates and by systematic measurement limitations [5].

Finally, recent probabilistic frameworks argue that modelling shot outcomes alone leaves a substantial part of the scoring process unexplained, motivating complementary approaches that explicitly account for the generation of shooting opportunities in addition to their conversion [17].

## 2.4. Beyond xG: methodological remedies and alternative performance indicators

A growing number of studies explicitly aim to overcome the limitations of xG by refining shot-based metrics and introducing complementary indicators that capture aspects of performance not observable through shots alone. Within the xG context, several contributions encourage more informed uses of xG for performance evaluation, particularly to reduce the influence of chance in low-scoring contexts and to support long-term evaluations rather than match-by-match narratives [3]. Methodological remedies have also been proposed to improve robustness and interpretability, for instance, through probabilistic corrections that account for player and position-specific effects and reduce estimation noise [20]. In parallel, recent work highlights that even refined xG estimates can be confounded when used to infer latent finishing skill, motivating debiasing-oriented analyses and careful treatment of modelling assumptions [5]. More fundamentally, probabilistic frameworks that go beyond modelling shot outcomes alone argue that a substantial component of the scoring process lies in the generation of shooting opportunities, suggesting extensions that explicitly model shot occurrences in addition to conversion probabilities [17].

In addition to remedies that remain focused on the shot level, the literature also proposes alternative performance indicators designed to represent broader match dynamics. Specifically, models based on statistics such as ball possession and territorial occupation during the match have been studied as predictors of match outcome and compared with xG-based approaches, demonstrating that integrating information on game control and ball progression can add explanatory power beyond shot quality alone [8]. Overall, these strands support a common conclusion: while xG provides a useful probabilistic description of shot quality, addressing its practical and conceptual limitations often requires strengthening the metric with more robust statistical formulations or integrating it with indicators that capture the non-shooting-related processes that drive match performance [3, 8].

Despite these advances, several gaps remain open that align closely with the aims of this study. First, although many works report improved xG pipelines, there is still a lack of a clear quantitative assessment of how far shot-based xG effectiveness can be pushed using publicly available data and carefully engineered spatial context. Second, existing implementations often remain difficult to reproduce and compare due to heterogeneous preprocessing choices, feature definitions, and evaluation protocols, leaving the field without a public, end-to-end operational reference for standardised xG construction and assessment. Third, while limitations of xG are frequently discussed, fewer studies provide a structured analysis of its intrinsic statistical and conceptual limitations and of the practical misinterpretations that follow from treating xG as a deterministic measure of match superiority. Finally, the growing body of works which try to go beyond xG metric highlights the need for complementary indicators, yet systematic comparisons that quantify how much additional explanatory power such models provide alongside xG remain limited.

This thesis addresses these points by developing a highly effective shot-level xG model enriched with engineered freeze-frame context (RQ1), releasing a transparent and reproducible modelling pipeline from data ingestion to evaluation (RQ2), critically analysing residual errors and common misuses of xG when moving from shot-level probabilities to match-level narratives (RQ3), and benchmarking event-based and combined models that complement xG with broader match information under consistent splitting and evaluation (RQ4).

# 3. Background

This chapter summarises the core concepts and notation required to understand the methods and results presented in the remainder of the thesis. We first introduce the main data modalities used in football analytics (event and positional data) and clarify the role of freeze-frame snapshots as a practical representation of shot-time spatial context. We then formalise expected goals (xG) as a probabilistic model of shot conversion and review the key families of shot descriptors used in xG modelling, from geometric features to contextual and spatial variables derived from player configurations. Finally, we outline the evaluation principles for probabilistic predictions and the basic ideas behind match-level performance modelling and model interpretability. These foundations provide a common vocabulary for the modelling pipelines developed in the subsequent chapters.

## 3.1. Football data: event data, positional data, and freeze-frames

Modern football analytics relies on two main data modalities. *Event data* represent the match as a sequence of discrete actions (e.g., passes, carries, shots, pressures), each associated with metadata such as timestamp, team and player identifiers, event type, outcome labels, and often a ball location on the pitch. This representation is compact, widely available across competitions, and well suited for constructing match summaries and shot-level models, but it captures only a partial view of the spatial configuration of players at each moment.

In contrast, *positional data* (also referred to as tracking data) provide time-continuous measurements of player and ball locations, typically sampled at high frequency. This modality enables a richer description of game context, including team shape, spacing, defensive pressure, and off-ball movement patterns, at the cost of higher storage requirements and more complex preprocessing [11]. Because full tracking data are often proprietary or not publicly accessible, some datasets provide intermediate representations that retain part of the spatial context without requiring full temporal trajectories.

A widely used compromise is the *freeze-frame* representation, where the locations of the goalkeeper and outfield players are recorded at the instant of a key event, most notably at shot time. Freeze-frames preserve the local spatial arrangement surrounding the shooter, making it possible to derive features related to defensive proximity, shot obstruction, and the effective shooting window, while remaining closer in structure and availability to event datasets. In this thesis, freeze-frame information derived from public StatsBomb data is used to augment event-based shot descriptors with engineered spatial context [21], bridging the gap between purely event-based xG models and fully tracking-based approaches.

## 3.2. Expected goals as a probabilistic model

Let $y \in \{0, 1\}$ denote the outcome of a shot, where $y = 1$ indicates a goal and $y = 0$ a non-goal. An xG model assigns to each shot described by a feature vector $x$ a probability

$$\mathrm{xG}(x) = P(y = 1 \mid x), \tag{1}$$

which can be interpreted as the estimated probability that the attempt results in a goal given the contextual information available at shot time, denoted with x in the equation. Under this formulation, xG is not a deterministic statement about whether a specific shot *should* score; rather, it is a probabilistic estimate whose meaning is inherently statistical.

The term *expected* reflects the fact that, over a sufficiently large collection of shots with comparable characteristics, the average observed conversion rate is expected to approach the predicted probability. In other words, if a set of similar shots are each assigned an xG value of 0.20, one would expect roughly 20% of them to result in goals in the long run. Deviations over short horizons, such as within a single match or a small number of attempts, are therefore not only possible but expected due to the random nature of shot outcomes.

At an aggregate level, summing shot probabilities across a set of attempts yields an estimate of the *expected number of goals*:

$$\mathrm{xG}_{\text{team}} = \sum_{i=1}^{n} \mathrm{xG}(x_i), \tag{2}$$

which corresponds to the expected value of the goal count under the modelling assumptions. This quantity is frequently used as a compact summary of chance quality and volume. However, it remains an expectation rather than a realised outcome, and its interpretation becomes less reliable as the number of shots $n$ decreases. These distinctions are central for both the evaluation of xG models at shot level and the correct interpretation of aggregated xG in applied analysis.

## 3.3. Feature representation for shot-based modelling

Shot-based models rely on a compact representation of each attempt through a set of predictors that encode *where* the shot was taken and *under which conditions* it occurred. In this thesis, features are grouped into three main categories: geometry, event context, and local defensive context.

### 3.3.1 Geometry

Geometric features describe the relationship between the shooting point and the target (the goal). They typically include distance-to-goal and angle-to-goal formulations, as well as derived quantities that capture the effective goal opening from the shooter's viewpoint (e.g., visible goal angle). These variables provide a first-order approximation of shot difficulty and often explain a large share of the variance in conversion probability.

### 3.3.2 Event context

Contextual variables encode attributes of the action that are not determined by location alone, such as body part, shot type (open play vs set piece), assist characteristics, or other categorical descriptors available in event logs. These features help distinguish attempts that are geometrically similar but arise from different tactical situations and execution constraints.

### 3.3.3 Local defensive context

Finally, spatially engineered variables aim to represent the immediate opposition environment around the shooter at the instant of the attempt. Typical examples include distances to the nearest defenders, counts of opponents within a given radius, goalkeeper proximity and relative positioning, and occlusion proxies that approximate how much of the goal is effectively available. By modelling this local configuration, the representation can separate shots that share similar geometry yet differ substantially in the degree of pressure or obstruction imposed by defenders.

## 3.4. Model families for xG

Given a feature representation $x$ for each shot, an xG model estimates a probability $\hat{p} = P(y = 1 \mid x)$ through a supervised classification approach. Several model families have been adopted in the literature and in applied pipelines, with different trade-offs between interpretability, flexibility, and computational cost.

### 3.4.1 Generalised linear models

A common baseline is logistic regression, which models the odds of scoring as a linear function of the input features. These models are simple, fast, and relatively interpretable, but they can struggle to capture non-linearities and high-order interactions unless extensive feature engineering is performed.

### 3.4.2 Tree-based models and ensemble methods

Decision trees and, in particular, ensembles such as random forests and gradient boosting machines are widely used for tabular football data because they naturally model non-linear effects, handle complex interactions, and accommodate heterogeneous feature types. Gradient boosting constructs an additive model by sequentially fitting weak learners to reduce the current error, resulting in a flexible predictor that often achieves strong performance on structured datasets [9].

### 3.4.3 Gradient boosting decision trees with LightGBM

In this thesis, the xG model is implemented using LightGBM, a highly efficient gradient boosting framework based on decision trees [10]. LightGBM is well suited to large-scale tabular learning settings and provides practical capabilities that align with the needs of xG modelling, including efficient training, regularisation controls, and native handling of categorical variables [1]. Empirically, gradient boosting methods are frequently competitive or superior to alternative learners on classification tasks involving structured features [7]. For these reasons, LightGBM was selected as the primary modelling approach to estimate shot conversion probabilities in the proposed xG pipeline.

## 3.5. Model interpretability with SHAP

As predictive models become more flexible, interpreting *why* a given prediction is produced becomes less straightforward. In football analytics, interpretability is particularly important because model outputs are often used to support decisions, scouting reports, or communication to non-technical stakeholders. In this thesis, model explanations are obtained using SHAP (SHapley Additive exPlanations), a framework that attributes a prediction to individual input features in a principled way [14].
SHAP is grounded in the concept of Shapley values from cooperative game theory. Given a model $f(\cdot)$ and an input instance $x$, the prediction $f(x)$ is expressed as an additive decomposition around a baseline value:

$$f(x) \approx \phi_0 + \sum_{j=1}^{p} \phi_j, \tag{3}$$

where $\phi_0$ is the expected model output under a chosen background distribution and each $\phi_j$ quantifies the contribution of feature $j$ to the prediction for that specific instance. Shapley values are defined as average marginal contributions across all possible feature coalitions, which gives SHAP desirable properties such as local accuracy (additivity), consistency, and a clear baseline reference [14].
In practice, SHAP enables both *local* and *global* interpretation. Local explanations describe which variables drive the prediction for an individual shot or match. Global analyses aggregate absolute SHAP values across many instances to rank the most influential features overall, and summary plots visualise how changes in feature values tend to increase or decrease the predicted probability. For tree-based models such as gradient boosting, efficient variants of SHAP exploit the structure of decision trees to compute explanations at scale, making it feasible to interpret large football datasets [14].

## 3.6. Evaluation of probabilistic predictions

Since xG models output probabilities rather than hard class labels, evaluation should assess both the *quality of the probability estimates* and, when needed, the behaviour induced by a chosen decision threshold. In this thesis, model effectiveness is primarily evaluated at shot level.

### 3.6.1 Proper scoring rules

Proper scoring rules quantify how well predicted probabilities match observed outcomes and encourage well-calibrated estimates. Two widely used examples are logarithmic loss (log-loss) and the Brier score. Log-loss penalises overconfident incorrect predictions more strongly, while the Brier score measures the mean squared error between predicted probabilities and binary outcomes. Both metrics directly evaluate the probabilistic meaning of xG, comparing the predicted likelihood of scoring to the realised shot outcome.

### 3.6.2 Discrimination metrics

Discrimination refers to the ability of the model to rank positive instances (goals) above negative instances (non-goals). The ROC–AUC measures ranking accuracy across all thresholds, while the precision–recall (PR) curve and its area (PR–AUC) are often more informative under class imbalance, as goals are relatively rare events.

### 3.6.3 Threshold-based metrics

For completeness, probabilistic predictions can also be converted into class labels using a probability threshold (e.g., 0.5) which represents the probability above which a shot is classified as a goal (class 1) by the model, enabling the computation of accuracy, precision, recall, and F1-score. Unlike proper scoring rules, these metrics depend on the selected threshold and therefore reflect a specific operating point rather than intrinsic probability quality. In the context of xG, they provide an additional perspective on how the model behaves when interpreted as a binary classifier, but they should be interpreted alongside probability-focused metrics.

## 3.7. Match-level outcome classification and evaluation metrics

Beyond shot-level modelling, this thesis also considers match-level prediction tasks where the target is the final outcome of a game. Match outcomes are represented as a three-class classification problem, with labels defined as loss ($-1$), draw ($0$), and win ($+1$) from the home team perspective. Each match is described through aggregated event-based indicators and, in the combined approach, additional team-level xG features.

For multi-class classification, the predicted outcome can be obtained either by selecting the class with maximum predicted probability (argmax) or by applying class-specific decision rules. Effectiveness is summarised using both overall and class-sensitive metrics. *Accuracy* measures the fraction of correctly classified matches. However, because draws are often less frequent and harder to predict, accuracy alone may hide unbalanced behaviour. For this reason, *macro-averaged F1-score* is used to weight classes equally: it is computed by first calculating precision and recall for each class and then averaging the resulting F1-scores across the three outcomes.
Class-wise precision and recall provide a more diagnostic view of model behaviour. Precision for a class quantifies how often predicted outcomes of that class are correct, while recall measures how often true outcomes of that class are successfully identified. Finally, the confusion matrix provides a compact summary of misclassification patterns by counting how many matches of each true class are assigned to each predicted class, revealing systematic confusions (e.g., draw vs win/loss) that may not be visible from aggregate metrics alone.

# 4. Proposed solution

## 4.1. Dataset

The publicly available StatsBomb Open Data release was used in this study [21]. The dataset provides detailed event-level information for professional football matches together with freeze-frame snapshots for every shot, covering multiple seasons of top-tier competitions, including UEFA Champions League, La Liga, Serie A, Bundesliga, Ligue 1, and Premier League. All matches were annotated using StatsBomb's standard event taxonomy, which includes precise timestamps, event types, player identifiers, body-part information, and spatial coordinates on a 120 × 80 meters pitch template.
Table 1 provides a detailed breakdown of the male competitions and seasons used in this study, together with the corresponding number of matches and shots.

Table 1: Competitions included in the study, aggregated by competition (summing matches and shots across all available editions)

| Competition | Season | Matches | Shots |
|---|---|---|---|
| La Liga | 18 | 868 | 21,210 |
| Premier League | 2 | 418 | 10,837 |
| Ligue 1 | 3 | 435 | 10,346 |
| Serie A | 2 | 381 | 10,033 |
| 1. Bundesliga | 2 | 340 | 8,747 |
| FIFA World Cup | 8 | 147 | 3,904 |
| Indian Super League | 1 | 115 | 3,095 |
| UEFA Euro | 2 | 102 | 2,629 |
| African Cup of Nations | 1 | 52 | 1,244 |
| NWSL | 1 | 36 | 1,041 |
| Copa America | 1 | 32 | 790 |
| Champions League | 18 | 18 | 594 |
| Major League Soccer | 1 | 6 | 146 |
| UEFA Europa League | 1 | 3 | 89 |
| Copa del Rey | 3 | 3 | 74 |
| Liga Profesional | 2 | 2 | 56 |
| North American League | 1 | 1 | 50 |
| FIFA U20 World Cup | 1 | 1 | 30 |
| Total | - | 2,960 | 74,915 |

### 4.1.1 Event Data

Each match contains a complete sequence of on-ball events such as passes, carries, pressures, duels, dribbles, interceptions, tackles, blocks, clearances, and shots. Every event includes the $(x, y)$ location on the pitch, the acting player, the event subtype (e.g. pass body part, shot technique) and contextual information such as whether the player was under pressure, the action occurred during a fast break or after a set piece. The event stream was temporally ordered, providing a high-level description of match dynamics.

### 4.1.2 Freeze-frame snapshots

For every shot, StatsBomb provides a freeze-frame capturing the positions of all 22 players and the ball at the moment the shot was taken. For each player, the dataset provides the $(x, y)$ coordinates, the role (position) and a flag that indicates whether the player is a opponent or a teammate of the shooter. Freeze-frame data has been widely used in recent contextual football models [2, 20] and is essential to approximate the spatial complexity and the context underlying shot outcomes.

### 4.1.3 Data preprocessing

All shot events available from men's competitions in the StatsBomb Open Data release were used, including open-play shots, set pieces, fast breaks, counterattacks, and penalties. No manual filtering based on shot type or contextual conditions was applied. Shot records were kept exactly as provided in the original dataset, preserving the full distribution of shooting situations.
Each shot carries a reference to its corresponding freeze-frame, which was accessed dynamically when computing spatial or positional features.
After loading the event data and linking shots to their associated freeze-frame metadata, the working dataset comprises *74,915 shots* across *2,960 professional matches*. These shot records serve as the foundation for all subsequent feature engineering and modelling steps.

### 4.1.4 Outcome definition

- Shot-level outcome (RQ1–RQ3): for xG modelling and its critical assessment, the target variable is binary and indicates whether a shot resulted in a goal ($y = 1$) or not ($y = 0$). Unless otherwise specified, model development, calibration, and interpretability analyses are performed at shot level.
- Match-level outcome (RQ4): to investigate whether xG can be complemented with broader indicators of performance, a second task is defined at the match level. Here, the target variable is the final match result in three classes (home win, draw, away win). Match-level features are obtained by aggregating the predicted shot probabilities for each team within a match (e.g., team-level expected scoring indicators) and using them, together with event-based variables, as inputs to the match-outcome prediction models.

## 4.2. Feature engineering

To approximate the typical accuracy level of contemporary xG models, an extensive set of spatial, contextual, match-state and freeze-frame features was constructed, enriching the variables already provided in the original StatsBomb dataset. The full feature set was designed to capture both the intrinsic characteristics of each shot and the spatial configuration of players at the moment of the attempt.

### 4.2.1 Event-level features (original dataset)

The base event data include information related to shot context and execution: period, timestamp (minute and second), play pattern, player position, shot technique, body part, whether the shot was taken under pressure, whether it was first-time, one-on-one, aerial, open-goal, followed a dribble, directly/indirectly from kick-off, whether was taken with the weak-foot and acrobatic-shot flags. These features form the foundational descriptors of shot type and context.

### 4.2.2 Spatial features (engineered)

Several geometric descriptors derived from pitch coordinates were incorporated to better capture shot difficulty. These include the shooter's $(x, y)$, location and goal-base angle. Additional kinematic descriptors derived from the key pass leading to the shot include pass length, speed, verticality, whether the pass entered the box, and whether it originated from a set piece. These spatial and directional features aim to quantify both the geometry of the shooting opportunity and the positional technical context in which it occurs.

### 4.2.3 Freeze-frame features (engineered)

Freeze-frame snapshots provide the positions of all players at the moment of the shot, enabling computation of defensive pressure and spatial congestion. From these frames were derived:
- the minimum distance to any defender ($d_{\min}$),
- counts of defenders within 1 m and 2 m of the shooter

- local pressure measures based on defender density
- goalkeeper positioning variables (offset from goal center and distance to the shooting line)
- shooter visible goal angle, shooter visible goal ratio (how much of the visible goal angle is blocked by opponents)
- shooter direction before taking the shot

These features capture defender proximity, goalkeeper positioning, and the overall spatial structure surrounding the shot. An example of a shot context is shown in Figure 1.

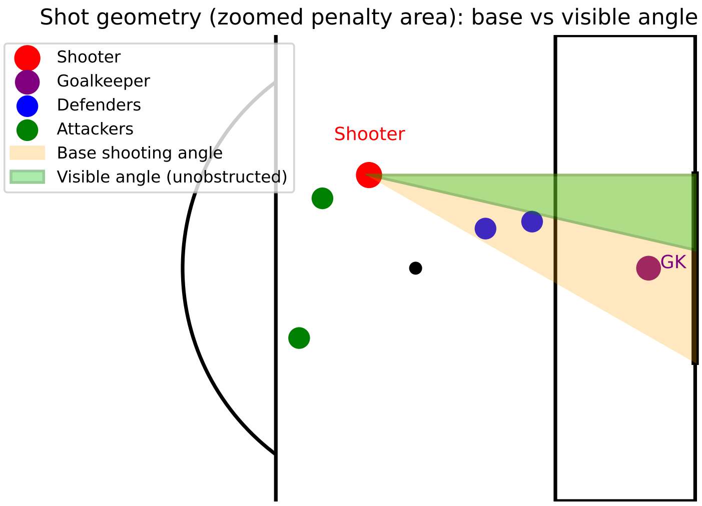



Figure 1: **Illustration of geometric shot features.** The diagram shows the relationship between the *base shooting angle* (orange) and the *visible angle* (green), computed from the shooter's location to the goal. The base angle corresponds to the angle subtended by the two goalposts, while the visible angle represents the unobstructed portion of this sector after accounting for defenders and the goalkeeper, who partially occlude the shooting cone. A wider visible angle typically indicates a clearer scoring opportunity, whereas defensive pressure reduces the effective space available to the shooter.

#### 4.2.4 Contextual and match-state features

Finally, match-level context was integrated including home/away indicator, scoreline for each team before the shot, score differential, bucketed score state, inferred play context, whether the shot was assisted, and cluster assignments describing the latent shot type distribution. These variables help encode behavioural and strategic variation related to match momentum, game state, and tactical patterns.
Together, these event-level, spatial, freeze-frame and contextual descriptors form a comprehensive feature set (Table 2) aimed at maximizing the representational capacity of xG models under realistic data constraints.

### 4.3. xG modelling framework

Shot success was modelled using an interpretable gradient-boosted decision tree (LightGBM). LightGBM was employed as the modelling algorithm due to its strong performance on structured tabular data [7], its ability to capture non–linear feature interactions [10], and its robustness in the presence of heterogeneous inputs such as spatial, contextual, and categorical variables [1]. Gradient boosting decision trees are a well–established choice for probabilistic classification and are widely used in domains with similar data characteristics [9]. LightGBM provides fast training, built–in regularization, and reliable probability estimates, making it a practical and methodologically sound option for constructing a reference xG model without requiring deep learning architectures [10].
The model predicts the probability that a shot results in a goal. Let $y \in \{0, 1\}$ denote the shot outcome (goal vs. no-goal).

Table 2: List of all features used in the xG model. Features are grouped into event-based (E), freeze-frame derived (F), engineered spatial metrics (S), contextual variables (C), and metadata (M).

| Feature | Type | Description |
|---|---|---|
| `match_id_x` | M | Match identifier (event table). |
| `competition_id_x` | M | Competition identifier. |
| `period` | E | Match period (first half, second half, etc.). |
| `minute`, `second` | E | Event timestamp within match. |
| `play_pattern_id`, `play_pattern_name` | E | Pattern of play (open play, set piece, etc.). |
| `position_id` | E | Player position at the shot moment. |
| `x`, `y` | E | Shot location on the pitch. |
| `shot_first_time` | E | Whether the shot is taken first-time. |
| `shot_technique_id` | E | Shot technique (normal, volley, etc.). |
| `shot_body_part_id` | E | Shooting body part. |
| `shot_type_id`, `shot_type_name` | E | Shot category (open play, free-kick, etc.). |
| `under_pressure` | E | Whether shooter was recorded under pressure. |
| `shot_aerial_won` | E | Aerial duel won prior to shot. |
| `shot_one_on_one` | E | One-on-one situation with goalkeeper. |
| `shot_open_goal` | E | Open goal opportunity. |
| `shot_follows_dribble` | E | Shot immediately following a dribble. |
| `shot_kick_off` | E | Shot taken directly from kickoff. |
| `preferred_foot_inferred` | S | Estimated preferred foot of shooter. |
| `p_left` | S | Probability shooter is left-footed. |
| `confidence` | S | Confidence of preferred-foot inference. |
| `is_weak_foot` | S | Whether the shot is taken with the weak foot. |
| `weak_x_distance` | S | Lateral deviation penalizing weak-foot shots. |
| `is_acrobatic` | S | Indicator for acrobatic shots. |
| `play_context` | S | Encodes tactical/phase context. |
| `goal_base_angle` | S | Geometric angle subtended by goal posts. |
| `goal_visible_angle` | S/F | Visible goal angle conditioned on defender positions. |
| `goal_visible_ratio` | S/F | Fraction of goal not obstructed by defenders. |
| `pressure_local` | F | Local defensive density around shooter. |
| `n_close_1m`, `n_close_2m` | F | Number of defenders within 1m/2m. |
| `dmin` | F | Minimum distance to nearest defender. |
| `gk_offset_from_center` | F | Lateral goalkeeper displacement relative to goal center. |
| `gk_dist_to_shotline` | F | Goalkeeper distance to predicted shot trajectory. |
| `kp_length` | E/S | Length of key pass preceding the shot. |
| `kp_speed` | S | Speed of key pass. |
| `kp_verticality` | S | Vertical component of key pass direction. |
| `kp_to_box` | E/S | Whether the key pass enters the penalty box. |
| `kp_from_set_piece` | E | Key pass originates from a set piece. |
| `is_assisted` | E | Whether the shot has an assisting pass. |
| `is_home_shot` | C | Whether the shooting team is playing at home. |
| `score_home_before`, `score_away_before` | C | Scoreline immediately before the shot. |
| `score_diff_shooter` | C | Score difference from shooter's perspective. |
| `score_state_shooter` | C | Encoded state (leading, drawing, losing). |
| `match_id_y` | M | Match identifier for freeze-frame table. |
| `score_diff_bucket` | C | Discretized score difference. |
| `shooter_direction` | S | Facing direction of the shooter. |
| `shooter_direction_source` | S | Method used to compute shooter direction. |
| `shot_cluster` | S | Cluster label from shot-position clustering. |
| `competition_id_y` | M | Competition identifier (freeze-frame table). |
| `competition_name` | M | Competition name. |
| `competition_gender` | M | Gender of competition. |

#### 4.3.1 Data splitting

A stratified five-fold partition of the dataset was defined at the shot level. For each rotation, three folds were used for model training, one fold was used for validation (hyperparameter selection), and one fold was held out for testing. This procedure was repeated five times, rotating the roles of the folds so that each fold served once as the test set.

#### 4.3.2 Hyperparameter selection

Within each rotation, hyperparameters were selected using the validation fold only. We tuned the LightGBM classifier hyperparameters using a successive-halving grid search (`HalvingGridSearchCV`) on the validation fold. The search space included the number of leaves, the minimum number of samples per leaf, the learning rate, and the fraction of features sampled per tree. Finally, the best configuration was selected according to the chosen validation metric (logarithmic loss).

#### 4.3.3 Model refitting and testing

After hyperparameter selection, the model was refitted on the combined training and validation folds using the selected configuration, and accuracy was evaluated exclusively on the held-out test fold. No information from the test fold was used during either hyperparameter selection or model fitting.

#### 4.3.4 Evaluation strategy

- Shot-level: model effectiveness was evaluated at the shot level using Brier score, logarithmic loss, and ROC-AUC, and threshold-based classification metrics (accuracy, precision, recall and F1-score) computed at a probability threshold of 0.5. In all cases, the predicted value represents the model-estimated *intrinsic scoring probability* (xG) for each shot, which is assessed against the observed binary outcome (goal vs. no goal). The proposed xG model accuracy was compared with the publicly available StatsBomb xG values provided in the dataset, testing also these values on the same test folds used for our new xG model. Also two non-informative baseline classifiers were used as baselines for the new model: the first baseline is a most-frequent classifier that always predicts the dominant outcome (no goal), while the second is a stratified classifier that samples predictions according to the empirical class distribution observed in the training data. All models were evaluated on identical data splits and metrics to ensure a fair comparison.

- Match-level: match-level evaluation was conducted by aggregating xG values for each team, in each match and using a simple Logistic Regression model. For each match, both teams' aggregated xGs (*Expected Goals*) were used as the only features for this simple model, that predicted every available match outcome as a three-class classification problem (Win home: 1, Draw: 0, Win Away: -1).

The objective of these stages was not to create the single best possible predictive model, but to approximate a realistic upper bound on xG metric accuracy using only publicly available data. Model interpretability was ensured through feature importance analysis and SHAP [14] value inspection. SHAP explains model predictions by decomposing it into a sum of feature contributions. For each observation, SHAP values provide individual (for each feature used by the model) explanations explaining how much each feature increases or decreases the predicted goal probability, providing fundamental results for the interpretability of the model and for understanding which are the most important variables that it relies on in its predictions.
We refer to the proposed model as *xG-SpatialAware Model*, as it incorporates engineered spatial features capturing the geometric relationship between the shooter, the goal, and surrounding defenders at the time of the shot.

### 4.4. Event-based match outcome model

To complement the limitations observed in traditional xG-based evaluations, an alternative, event-based model was introduced to assess team performance at the match level. In this setting, the prediction target is the final match outcome in three classes, encoded as $\{-1, 0, 1\}$ (away win/loss for the home team, draw, and home win, respectively). Unlike xG, which estimates the probability of an individual shot becoming a goal, this model aggregates a broad set of in-game actions and constructs a predictive representation of match dynamics [8]. The goal of the metric is not to forecast goals directly, but rather to quantify the extent to which a team's overall event production aligns with the final match outcome.

#### 4.4.1 Input features

The model uses aggregated event descriptors extracted from the full match dataset, including counts and rates of passes, shots, dribbles, recoveries, defensive pressures, progressive passes, entries into advanced zones, interceptions, clearances, goalkeeper saves, and other on-ball actions. These features (Table 3) capture both attacking and defensive behaviour and reflect the overall flow and balance of match events. All variables were computed separately for the home and away team, allowing the model to encode relative dominance on a per-match basis.

Table 3: Features used in the event-based performance model. Unless otherwise noted, each variable was computed separately for the home and away team. Features summarise passing activity, shooting volume, defensive actions, recoveries, pressures, progression, possession dynamics, and pre-match priors.

| Feature | Type | Description |
|---|---|---|
| `index` | M | Match index for each team. |
| `competition_id` | M | Competition identifier. |
| `season_id` | M | Season identifier. |
| `n_pass` | E | Total attempted passes. |
| `n_pass_ok` | E | Successful passes. |
| `pass_acc` | S | Pass accuracy. |
| `n_prog_pass` | E/S | Progressive passes. |
| `prog_pass` | S | Distance-based progressive passing measure. |
| `n_cross` | E | Crosses attempted. |
| `n_deep_pass` | E | Passes into deep areas. |
| `n_pass_final_third` | E | Passes into the final third. |
| `pass_final_third_share` | S | Share of passes targeting final third. |
| `n_box_entry_pass` | E | Passes entering the penalty box. |
| `deep_pass_rate` | S | Rate of deep passes. |
| `n_carry` | E | Total carries. |
| `n_prog_carry` | E/S | Progressive carries. |
| `prog_carry` | S | Progressive carrying distance. |
| `dangerous_carry` | S | Carries entering threatening areas. |
| `n_box_entry_carry` | E | Carries entering the penalty box. |
| `n_shot` | E | Total shots. |
| `n_shot_ot` | E | Shots on target. |
| `n_shot_box` | E | Shots inside the box. |
| `n_shot_blk` | E | Blocked shots. |
| `sot_rate` | S | Proportion of shots on target. |
| `shot_box_rate` | S | Proportion of shots taken inside the box. |
| `n_keypass` | E | Total key passes. |
| `keypass_rate` | S | Key passes per possession. |
| `n_pressure` | E | Total pressures. |
| `n_pressure_high` | E | Pressures in high defensive zones. |
| `n_pressure_regain` | E | Pressures leading to regain. |

*Continued on next page*

Table 3 (continued)

| Feature | Type | Description |
| --- | --- | --- |
| n_pressure_high_regain | E | High pressures leading to regain. |
| counterpress_won | S | Successful counterpressing actions. |
| n_interception | E | Interception attempts. |
| n_interception_won | E | Successful interceptions. |
| n_recovery | E | Ball recoveries. |
| recovery_high | S | Recoveries in high zones. |
| recovery_mid | S | Recoveries in midfield zones. |
| recovery_low | S | Recoveries in deep zones. |
| n_clearance | E | Clearances. |
| n_duel | E | Duel attempts. |
| n_duel_won | E | Duels won. |
| duel_win_rate | S | Duel win percentage. |
| n_dribble | E | Dribble attempts. |
| n_dribble_w | E | Successful dribbles. |
| dribble_win_rate | S | Dribble success rate. |
| n_foul_committed | E | Fouls committed. |
| n_yellow | E | Yellow cards. |
| n_red | E | Red cards. |
| cards_per_foul | S | Cards per foul ratio. |
| n_touch | E | Total touches. |
| n_touch_final_third | E | Touches in the final third. |
| touch_final_third_share | S | Share of touches in final third. |
| deep_touches | E | Touches in deep attacking areas. |
| poss_seconds | S | Total time in possession (seconds). |
| field_tilt | S | Share of final-third actions controlled. |
| attacking_share | S | Proportion of attacking actions. |
| verticality | S | Verticality of passes and carries. |
| switches | E/S | Switches of play. |
| n_counter | E | Counterattacks initiated. |
| n_counter_shot | E | Counterattacks ending with a shot. |

### 4.4.2 Model formulation

A multiclass classifier was trained to predict the full-time match result from the home team's perspective, using three outcome labels: win (+1), draw (0), and loss (−1). The classifier was implemented using gradient boosting on decision trees (LightGBM), with hyperparameters tuned via five-fold cross-validation. No shot-level information or xG estimates were used at any stage in this stage.

### 4.4.3 Evaluation protocol

Model effectiveness was evaluated on different matches, using different classification metrics to assess the accuracy on each outcome-class (home win, draw or away win). Metrics include match-level accuracy, macro F1-score, per-class precision/recall, and confusion matrices.

### 4.4.4 Model interpretation

This event-based model does not aim to replace xG, but rather to provide a complementary perspective on team performance. Whereas xG captures the quality of shooting opportunities, the event-based metric reflects a broader conception of match dominance, incorporating the cumulative effect of possession, progression, chance creation and defensive organization impact.

## 4.5. Combined model

### 4.5.1 Model formulation

To assess the degree to which shot-level information complements aggregate match events, a combined model was constructed in which the team-level xG values produced by the reference xG-SpatialAware model were added as features to the event-based framework. This integration allows the model to incorporate both overall match behaviours and the quality of the scoring opportunities created, providing a more complete representation of team performance for match–outcome prediction.
Figure 2 shows a schematic overview of the modelling pipeline, explaining the data used at each step and the resulting models.

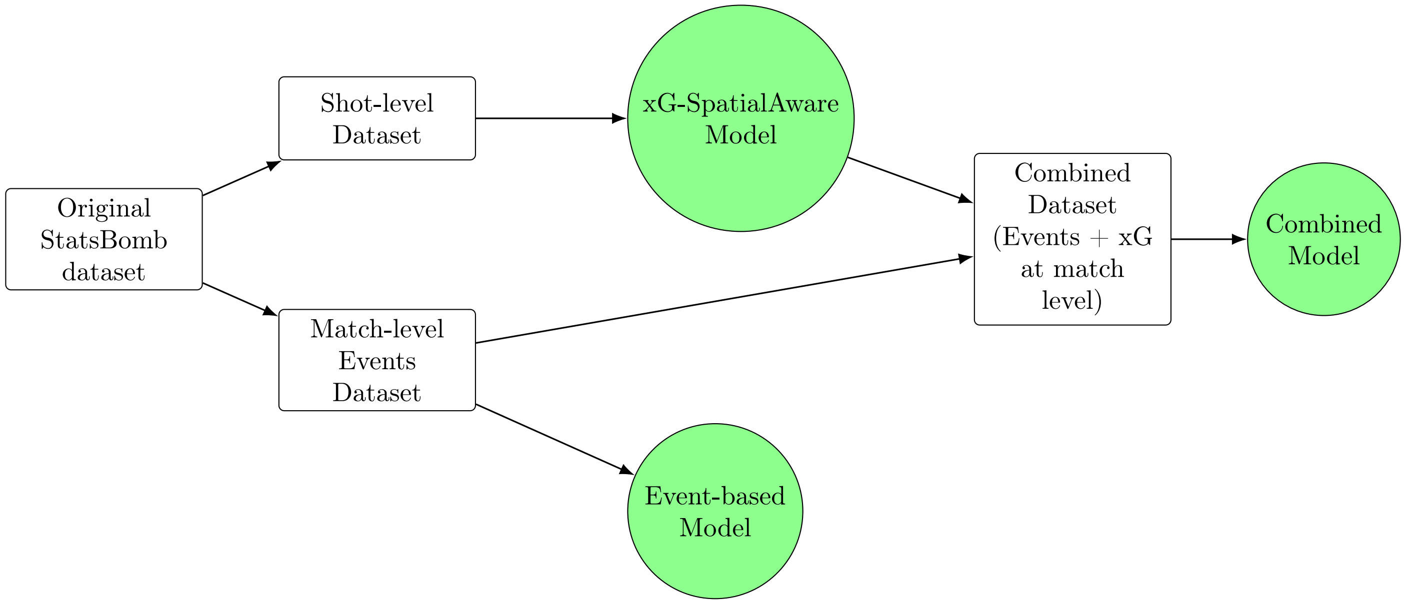



Figure 2: Schematic overview of the modelling pipeline. The xG-SpatialAware model uses shot-level event data and freeze-frame spatial information; the event-based model aggregates match-level actions; and the combined model integrates both sources to improve match outcome prediction.

### 4.5.2 Evaluation protocol

Model effectiveness was evaluated on different matches, using different classification metrics to assess the accuracy on each outcome-class (home win, draw or home loss/away win). Metrics include match-level accuracy, macro F1-score, per-class precision/recall, and confusion matrices.

## 4.6. Final evaluation

The three proposed models are finally compared in terms of match-outcome prediction accuracy. For each match in the dataset, each model produces a predicted outcome (home win, draw, or away win), and standard classification metrics are computed by comparing these predictions against the observed results.
In addition to match-outcome classification, a final evaluation was conducted at the team–match level to assess the ability of the models to predict realised points. Concretely, for every match, each of the three models outputs a predicted outcome; this predicted outcome is then mapped to expected match points for each team according to the standard scoring rule (3 points for a win, 1 for a draw, 0 for a loss). Predicted and observed points are accumulated match by match over the full dataset, yielding, for each team, a final total of predicted points and a final total of realised points. Model performance is then assessed by comparing these team-level totals across all teams once all matches have been processed.

Model accuracy was quantified using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), computed between predicted and realized points. Errors were first calculated globally over the full dataset to obtain results of the global effectiveness of the models in terms of point predictions. Then, once all the expected points were predicted, a further analysis was conducted at competition-level analysing the MSE and RMSE for each competition, in order to assess model behaviour across different competitive contexts and seasons.
This evaluation complements the match-level classification analysis by focusing on longer-horizon performance consistency and point estimation accuracy, providing an additional perspective on the practical reliability of the proposed models.

# 5. Experimental results

## 5.1. xG-SpatialAware model effectiveness

The proposed xG-SpatialAware model shows consistently strong predictive effectiveness across all cross-validation splits. Using a five-fold stratified cross-validation scheme, the model achieves average values of 0.2624 in log-loss, 0.0751 in Brier score, and 0.8292 in ROC–AUC, with limited variability across folds. These results indicate that the model captures a substantial portion of the available signal in shot outcomes, despite the intrinsic complexity and high randomness associated with goal-scoring events [3].
Model interpretability was assessed using SHAP [14] to quantify the contribution of individual features to xG predictions. Figure 3 reports the SHAP summary plot for the xG-SpatialAware model, providing insight into how individual shot-level features contribute to goal probability estimates. As expected [2], geometric descriptors dominate the explanation. The base shooting angle and the visible goal angle are two of the most influential variables, with larger angles strongly increasing the predicted likelihood of scoring. Shot location along the longitudinal axis further reinforces this effect, meaning that the closer is the shooter to the goal, the better are the chance to score.
Freeze-frame derived variables related to defensive pressure and goalkeeper positioning also exhibit consistent directional contributions, indicating that spatial congestion and goalkeeper misalignment affect shot quality estimates. Conversely, contextual and match-state features play a minor role, confirming that xG primarily encodes instantaneous shot conditions rather than broader match dynamics.

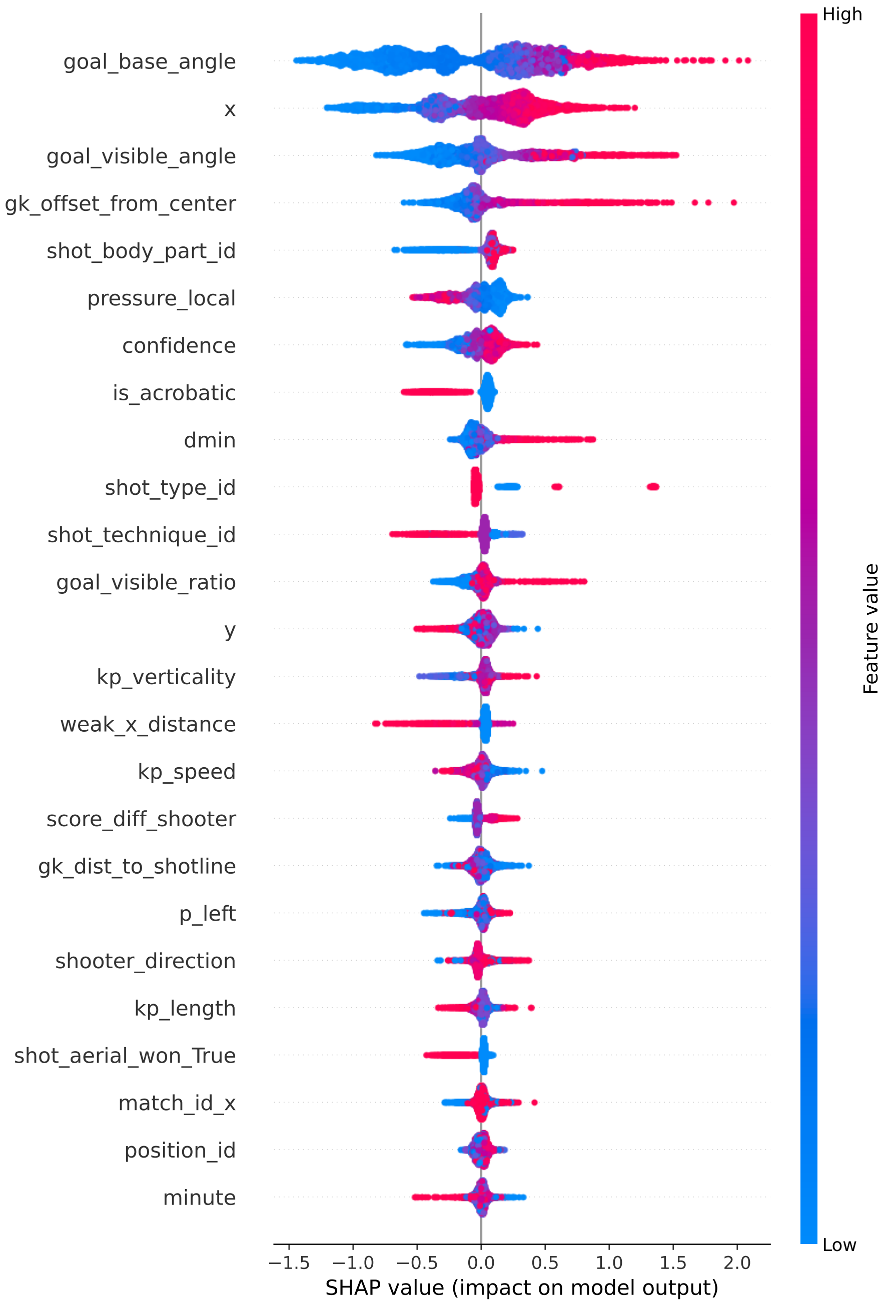



Figure 3: SHAP summary plot showing the top 15 features contributing to xG predictions. Positive SHAP values increase the predicted goal probability, while negative values decrease it.

**Comparison with public and state-of-the-art xG models**

Table 4 summarises the probabilistic accuracy of the proposed xG-SpatialAware model alongside the publicly available StatsBomb values and the ranges typically reported in recent literature. On the same five-fold cross-validation splits, the StatsBomb baseline reaches an ROC–AUC of 0.824 and a PR–AUC of 0.473, with a log-loss of 0.267 and a Brier score of 0.076. The precision, recall, and F1-score for the goal class (0.71, 0.23, 0.35) illustrate the inherent difficulty of predicting rare scoring events, and are consistent with expected behaviour for open-data xG models.

Our model achieves slightly higher effectiveness across all metrics (ROC–AUC 0.8292, PR–AUC 0.493, log-loss 0.2624, Brier score 0.0751), with goal-class precision, recall, and F1-score (0.700, 0.251, 0.369) aligning with the strongest results reported in the literature. These improvements do not stem from a single methodological choice but from the cumulative contribution of multiple feature additions, including spatial descriptors and contextual match variables derived from freeze-frame information.

It is important to note that published xG models vary substantially in dataset composition, feature

availability, temporal coverage, and modelling objectives. For this reason, the values reported here should not be interpreted as direct benchmarking. Rather, they position our model within the upper range of what is typically achievable under open-data constraints and support the conclusion that the proposed pipeline approximates a realistic performance ceiling for xG modelling.
A further result reinforces the structural limitations of xG as a standalone measure of team performance. In our experiments, the team generating the higher aggregate xG wins the match only 61% of the time. This demonstrates that, despite strong probabilistic calibration at the shot level, xG does not reliably translate into outcome-level predictive power. Match results depend on a wide range of behavioural and tactical factors that xG cannot capture, highlighting the need for complementary metrics, developed in the following sections, to obtain a more complete representation of team performance.

Table 4: Comparison of predictive accuracy between the proposed xG-SpatialAware model, the publicly available StatsBomb xG values [21], two dummy baselines (most-frequent and stratified), and representative performance ranges reported in recent literature [2, 4, 6, 15]. Results are reported as mean ± standard deviation over five stratified folds. Precision, recall, and F1-score refer to the positive class (goals).

| Model | ROC–AUC | PR–AUC | LogLoss | Brier | Precision (Goal) | Recall (Goal) | F1 (Goal) |
|---|---|---|---|---|---|---|---|
| **xG-SpatialAware** (ours) | $0.8308 \pm 0.0059$ | $0.4889 \pm 0.0086$ | $0.2616 \pm 0.0030$ | $0.0750 \pm 0.0007$ | $0.6970 \pm 0.0055$ | $0.2520 \pm 0.0090$ | $0.3700 \pm 0.0097$ |
| StatsBomb public xG | $0.8277 \pm 0.0067$ | $0.4784 \pm 0.0049$ | $0.2642 \pm 0.0027$ | $0.0754 \pm 0.0005$ | $0.7121 \pm 0.0161$ | $0.2342 \pm 0.0068$ | $0.3525 \pm 0.0092$ |
| Dummy (most-frequent) | $0.5000 \pm 0.0000$ | $0.1106 \pm 0.0000$ | $3.9871 \pm 0.0013$ | $0.1106 \pm 0.0000$ | $0.0000 \pm 0.0000$ | $0.0000 \pm 0.0000$ | $0.0000 \pm 0.0000$ |
| Dummy (stratified) | $0.5031 \pm 0.0024$ | $0.1113 \pm 0.0005$ | $7.0009 \pm 0.0336$ | $0.1942 \pm 0.0009$ | $0.1162 \pm 0.0043$ | $0.1144 \pm 0.0042$ | $0.1153 \pm 0.0042$ |
| Literature (top range) | 0.80–0.84 | 0.32–0.43 | 0.260–0.270 | 0.073–0.082 | 0.30–0.74 | 0.24–0.67 | 0.35–0.44 |

## 5.2. Event-based model effectiveness

The event-based model exhibits higher match-level predictive accuracy than the xG-based approach, despite relying solely on aggregated on-ball actions rather than shot-quality estimates. To obtain match-level predictions from xG-SpatialAware model, shot probabilities were first aggregated into team-level xG features for each match and then fed into a simple multinomial logistic regression model to predict the three-class outcome (win/draw/loss). Across the five cross-validation splits, the event-based model reaches an average accuracy of 0.657, compared to 0.623 for the xG-SpatialAware model baseline, indicating a better global effectiveness across win, draw, and loss predictions.
Confusion matrices (Table 4) reveal systematic differences between the two approaches. The event-based model shows a marked improvement in identifying both home wins and home losses, reducing misclassifications in the strongest outcome classes. Draws remain challenging for both models, consistent with their intrinsic unpredictability and limited representational signal.

Figure 4: Confusion matrices for the three-class outcome prediction task (rows=true, cols=pred; class order $\{HomeLoss(-1), Draw(0), HomeWin(1)\}$). Diagonal entries (bold) indicate correct predictions.

(a) xG-SpatialAware Model

| True \Pred | -1 | 0 | 1 |
|---|---|---|---|
| **-1** | **631** | 122 | 176 |
| **0** | 240 | **161** | 306 |
| **1** | 139 | 132 | **1051** |

(b) Event-based

| True \Pred | -1 | 0 | 1 |
|---|---|---|---|
| **-1** | **709** | 69 | 151 |
| **0** | 279 | **70** | 358 |
| **1** | 111 | 48 | **1163** |

The xG-SpatialAware model tends to overpredict outcomes aligned with shooting dominance, capturing scenarios where one team produces substantially higher shot quality. By contrast, the event-based model incorporates broader patterns of match control, benefiting from features related to ball progression, pressure, defensive actions, and goalkeeper activity. This allows it to correctly identify matches where the final outcome reflects cumulative territorial or tactical superiority rather than only shot

quality.
Overall, these findings suggest that event-based match-level representations complement xG by capturing dimensions of performance that are orthogonal to shot quality. While the event-based model does not evaluate scoring probability directly, it demonstrates a stronger alignment with final match outcomes, indicating that broader match dynamics play a crucial role in determining results and should not be overlooked in analytical workflows.

### 5.3. Combined xG-SpatialAware + Event-based model

To evaluate whether shot-level information provides complementary signal to aggregate match events, the event-based model was extended by including the team-level xG values obtained from the proposed xG-SpatialAware model. This combined approach consistently outperforms both individual models across all evaluation metrics.
Table 6 reports the classification metrics for match outcome prediction of all three models. The combined model achieves an overall accuracy of 0.676, compared to 0.657 for the event-based model and 0.623 for the xG-SpatialAware model. Macro-averaged F1-score also increases from 0.544 (event-based) and 0.558 (xG-SpatialAware) to 0.581 in the combined setting, indicating a more balanced accuracy across win, draw, and loss outcomes. Improvements were observed across all three outcome classes: the F1-score for home wins rises to 0.80, for draws to 0.22, and for away wins to 0.72.
The combined model's confusion matrix (Table 5) shows more improvements in identifying both home wins and home losses, reducing misclassifications in the strongest outcome classes. Draws recognition improves compared to the event-based model, but they remain highly challenging, consistent with their intrinsic unpredictability and limited representational signal.

Table 5: Combined model's confusion matrix for the three-class outcome prediction task (rows=true, cols=pred; class order $\{-1, 0, 1\}$). Diagonal entries (bold) indicate correct predictions.

| **True \Pred** | **-1** | **0** | **1** |
|---|---|---|---|
| **-1** | **727** | 78 | 124 |
| **0** | 279 | **110** | 318 |
| **1** | 88 | 70 | **1164** |

Table 6: Match outcome prediction accuracy for the xG-SpatialAware model, the event-based model, the combined model integrating team-level xG features, and a dummy baseline (stratified). Metrics are reported for the three outcome classes (loss = -1, draw = 0, win = 1). Dummy model is evaluated out-of-fold using GroupKFold.

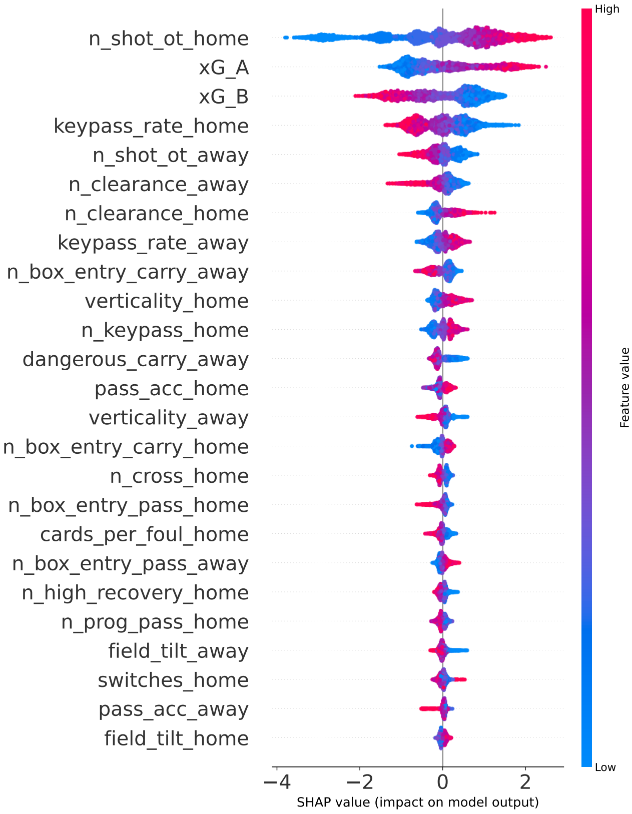
| Model | Accuracy | F1 (macro) | Prec L | Prec D | Prec W | Rec L | Rec D | Rec W |
|---|---|---|---|---|---|---|---|---|
| xG-SpatialAware model | 0.623 | 0.558 | 0.62 | 0.39 | 0.69 | 0.68 | 0.23 | 0.80 |
| Event-based | 0.657 | 0.544 | 0.65 | 0.37 | 0.70 | 0.76 | 0.10 | 0.88 |
| Combined xG + Event | **0.676** | **0.581** | **0.66** | **0.43** | **0.72** | **0.78** | 0.16 | **0.88** |
| Dummy (stratified) | $0.349 \pm 0.0120$ | $0.316 \pm 0.0119$ | $0.313 \pm 0.0220$ | $0.194 \pm 0.0329$ | $0.442 \pm 0.0329$ | $0.323 \pm 0.0155$ | $0.176 \pm 0.0274$ | $0.452 \pm 0.0154$ |

These results indicate that xG values encode relevant information about shot quality that is not captured by aggregated event descriptors alone, while the event-based model incorporates broader match dynamics and territorial patterns absent from xG. Their integration provides the most accurate representation of team performance for outcome prediction, confirming that the two perspectives are not redundant but instead reflect distinct and synergistic dimensions of match play.
To interpret the contribution of match-level variables to outcome predictions, the combined model was analysed using SHAP. Figures 5a–5c show SHAP summary plots for the three outcome classes (win, draw, and loss).

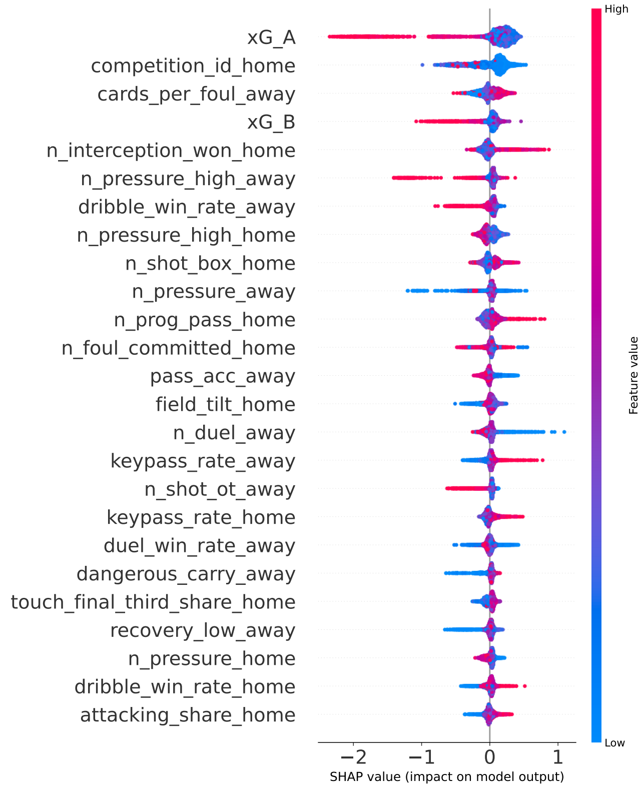
- For the *win* class, offensive production indicators for the home team (including shots on target, total xG, key pass rates, and box entries) give strong positive contributions, while analogous away-

team variables reduce win probability. Defensive actions by the away team further contribute negatively.
- The *draw* class exhibits a more balanced pattern, with pressing intensity, interceptions, and possession-related features playing a comparatively larger role, reflecting the difficulty of distinguishing evenly matched games.
- For the *loss* class, away-team attacking output and territorial control dominate the explanation, whereas home-team shooting and progression features mitigate loss probability. This mirrors the patterns observed for wins, revealing the symmetric structure of the event-based representation.

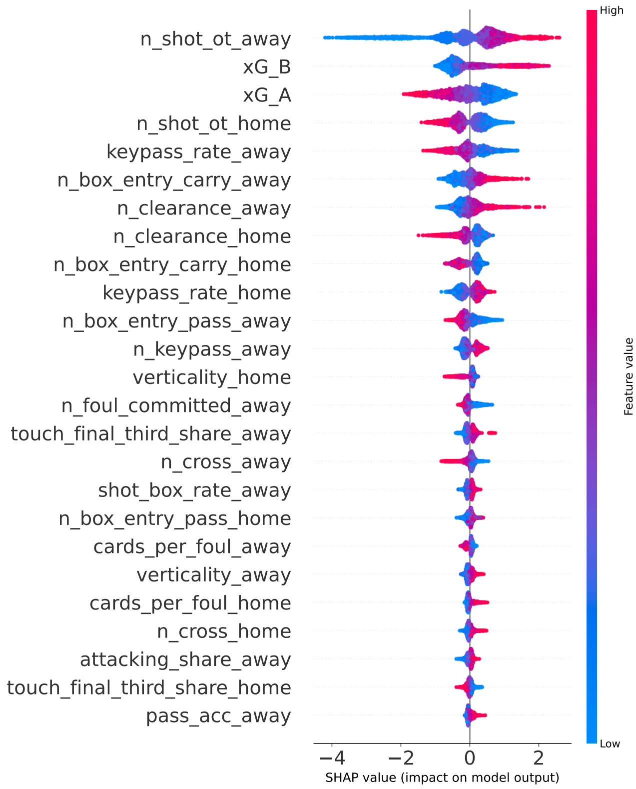
Overall, the SHAP analysis confirms that the event-based and the combined model captures cumulative attacking, defensive, and territorial behaviours over the full match, providing information that is structurally complementary to shot-based xG estimates.




(a) Home Win (+1)




(b) Draw (0)




(c) Home Loss (-1)

Figure 5: SHAP summary plots for the combined match outcome model. Each panel reports feature contributions for one outcome class: (a) win, (b) draw, and (c) loss. Positive SHAP values indicate an increased predicted probability for the corresponding class, while negative values indicate a decrease. Features are ordered by mean absolute SHAP value, highlighting the most influential variables driving match outcome predictions.

### 5.4. Long-term effectiveness and points prediction

To assess the ability of the proposed models to capture long-term team performance, expected points were compared against observed points at the team-level across the full dataset. For each match, each model produces a predicted outcome (win/draw/loss); this outcome is then mapped to predicted points using the standard scoring rule (3 points for a win, 1 for a draw, 0 for a loss) for both teams. This yields, for every team instance, a predicted points value that can be directly compared to the corresponding realised points. Effectiveness was evaluated using mean absolute error (MAE) and root mean squared error (RMSE), computed both globally across all competitions and separately for each competition.

At the global level, the event-based model substantially outperforms the xG-SpatialAware approach, reducing RMSE from 1.283 to 1.169 points per match. When team-level xG features are integrated into the event-based representation, the combined model achieves a further improvement, reaching the lowest global error (RMSE = 1.103), indicating a more accurate reconstruction of realised points. All results are reported in Table 7.

Table 7: Global error metrics for expected points prediction. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are reported for the xG-SpatialAware model, the event-based model, and the combined xG + event-based model. Lower values indicate better accuracy.

| Model | MAE | RMSE |
|---|---|---|
| xG-SpatialAware model | 0.725 | 1.283 |
| Event-based | 0.626 | 1.169 |
| Combined (xG + Event) | **0.581** | **1.103** |

Figure 6 compares the RMSE of predicted points across competitions with at least 50 matches for the xG-SpatialAware model, the event-based model, and the combined approach.

Table 8 shows the RMSE compared to real total points for each competition. Competitions where all three models have a high error percentage are those with a smaller number of matches available in the dataset, which is therefore not a large enough sample to reflect true models effectiveness.

Across almost all competitions, the xG-SpatialAware model exhibits the highest RMSE, indicating limited ability to infer long-term effectiveness trends from shot quality alone. The event-based model consistently reduces prediction error, highlighting the importance of aggregated match behaviours, such as territorial control, defensive actions, and progression patterns, in explaining seasonal outcomes.

The combined model further improves or stabilises accuracy in most competitions, achieving the lowest or near-lowest RMSE in the majority of cases. This effect is particularly evident in major domestic leagues and international tournaments, where the higher number of available matches allows more accurate estimates of accumulated points.

While the magnitude of improvement varies across competitions, the overall trend is consistent: incorporating xG into an event-based framework provides additional explanatory power beyond either component in isolation. These results suggest that long-term effectiveness is better captured by models that jointly account for chance quality and the underlying match processes that generate those chances.

The violin plot 7 summarizes the distribution of point-prediction error (RMSE) across random seeds, pooling all competitions with at least 50 matches.

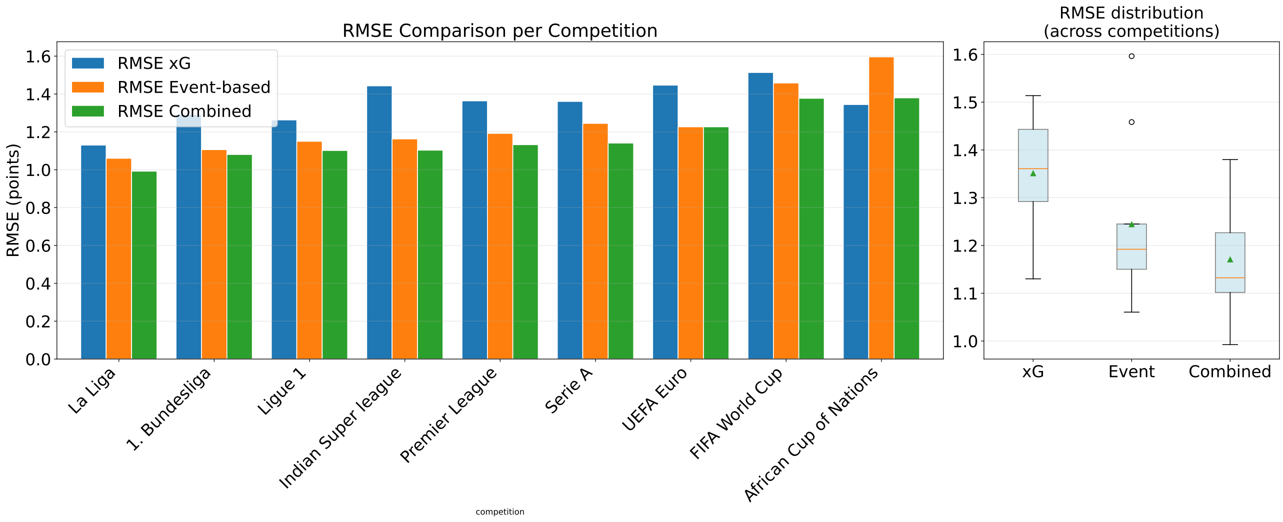



Figure 6: **Competition-level RMSE of predicted points.** Root mean squared error (RMSE) of predicted points per competition (with at leat 50 matches in the dataset) for the xG-SpatialAware model, the event-based model, and the combined approach. Lower values indicate better long-term estimation effectiveness.

Table 8: RMSE normalised by total real points per competition (percentage). Lower is better.

| Competition | Total points | RMSE_xG (%) | RMSE_Event (%) | RMSE_Combined (%) |
|---|---|---|---|---|
| FIFA U20 World Cup | 3 | 0.00 | 0.00 | 0.00 |
| Liga Profesional | 6 | 0.00 | 0.00 | 0.00 |
| North American League | 3 | 0.00 | 0.00 | 0.00 |
| La Liga | 2426 | 0.05 | 0.04 | 0.04 |
| Ligue 1 | 1188 | 0.11 | 0.10 | 0.09 |
| Premier League | 1134 | 0.12 | 0.11 | 0.10 |
| Serie A | 1044 | 0.13 | 0.12 | 0.11 |
| 1. Bundesliga | 939 | 0.14 | 0.12 | 0.12 |
| FIFA World Cup | 415 | 0.36 | 0.35 | 0.33 |
| Indian Super league | 316 | 0.46 | 0.37 | 0.35 |
| UEFA Euro | 275 | 0.53 | 0.45 | 0.45 |
| NWSL | 100 | 1.14 | 1.14 | 0.97 |
| African Cup of Nations | 138 | 0.97 | 1.16 | 1.00 |
| Copa America | 91 | 1.26 | 1.39 | 1.49 |
| Champions League | 54 | 2.66 | 2.27 | 1.85 |
| Major League Soccer | 17 | 3.80 | 3.80 | 3.80 |
| Copa del Rey | 9 | 19.25 | 0.00 | 10.14 |
| UEFA Europa League | 9 | 0.00 | 19.25 | 19.25 |

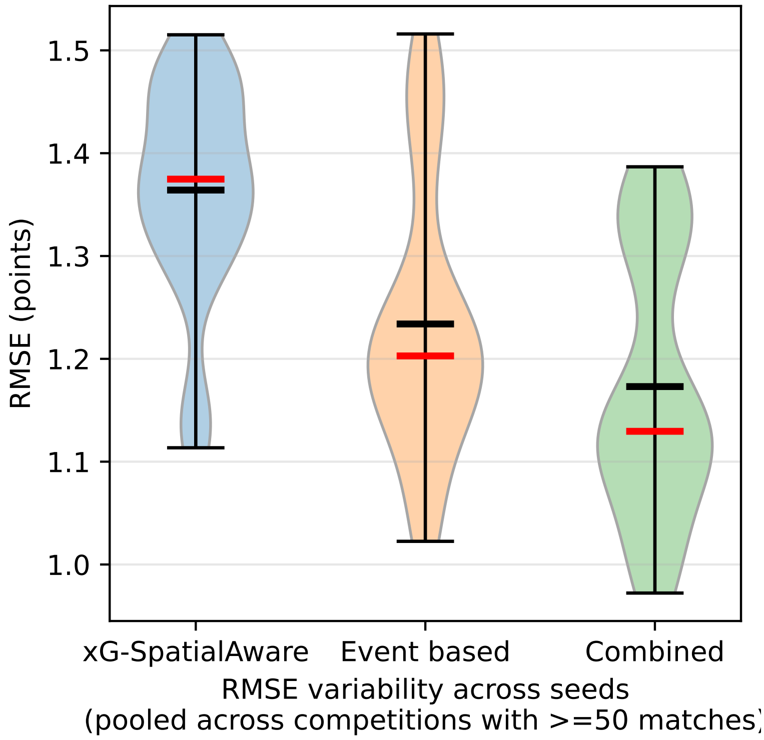



Figure 7: **Violin plots of team–match point-prediction error (RMSE) across random seeds, pooled over competitions with at least 50 matches.** Each violin shows the distribution of RMSE values obtained by repeating model training with different random initialisations for the xG-SpatialAware, Event-based, and Combined approaches. Lower values indicate better long-term estimation effectiveness.

# 6. Conclusions

## 6.1. Discussion

This study was guided by four research questions: improving xG predictive effectiveness, providing a publicly accessible and reproducible operational framework, clarifying the metric's strengths and intrinsic limitations, and assessing complementary modelling approaches where shot-based xG alone is insufficient.
The following subsections address each question in turn, summarising the key insights emerging from our analyses.

### 6.1.1 RQ1: How can xG models be improved using publicly available data?

The results show that xG effectiveness can be improved in a consistent way when the model is enriched with engineered spatial context computed from freeze-frame information.
The xG-SpatialAware model is based on features that describe the geometry of the shot and the local defensive pressure, allowing for a more precise description of the shooting context. In this way, the model can better distinguish shots that are superficially similar in location but differ substantially in how the opponents are positioned, how much pressure they put on the shooter and, above all, how much angle of the goal they leave free for the shooter, a variable which is one of the most important for the model after the shooter's position.
Overall, these design choices improved the predictive accuracy of the xG metric and positioned xG-SpatialAware among the strongest results reported in the literature for public-data xG modelling (Table 4). This supports the answer to RQ1: refining how the shooting context is represented through spatially informed features can yield measurable gains in xG model accuracy, providing a more reliable

probabilistic description of shot quality for practical football analysis.

### 6.1.2 RQ2: Can a public and accessible operational framework be defined to standardise xG construction and evaluation?

Beyond the empirical results, this study contributes to addressing the broader issue of the lack of an operationally consistent framework for xG modelling [23]. Current xG implementations differ widely across data providers and research groups, leading to substantial variation in how the metric is computed and how its values are interpreted [23]. While establishing a universal standard is beyond the scope of this work, the methodology presented here offers a step toward a more transparent and reproducible operational pipeline. By explicitly defining all stages of the modelling process (data preparation, feature engineering, freeze-frame interpretation, cross-validation strategy, and model evaluation) is provided a clear reference workflow that clarifies how specific modelling choices influence xG estimates. This level of transparency helps reduce the ambiguity that often surrounds xG construction and contributes to the ongoing effort to develop more unified methodological conventions for the field. In addition, releasing the full implementation as publicly accessible code (from data ingestion and preprocessing to feature computation, model training, and final evaluation) strengthens the answer to RQ2. Making the entire pipeline available ensures that it can be inspected, reproduced, and extended by other researchers and practitioners, providing a performant and accessible reference workflow for xG construction.

### 6.1.3 RQ3 : What are xG's intrinsic statistical and conceptual limitations, and how is it commonly misused or misinterpreted in practice?

This study provides a comprehensive reassessment of the Expected Goals (xG) framework and its role in modern football analytics. By constructing a reference xG model using publicly available event and freeze-frame data, it was shown that even highly engineered shot-probability estimators exhibit structural limitations that are intrinsic to the nature of shot outcomes [17]. These limitations persist regardless of model specification and arise from the high variance inherent in finishing actions [3], the influence of defender and goalkeeper behaviour, and the difficulty of fully representing spatial interactions from the available positional information, which remains sensitive to data limitations and unobserved factors [2, 20].
A further implication of our findings concerns the current use of xG in applied football analysis. In contemporary media and even in professional settings, match dominance is often inferred directly from the comparison of aggregate xG values, implicitly assuming that producing more xG should correspond to a higher probability of winning [22]. Our results show that this interpretation is fundamentally flawed: higher xG does not necessarily translate into match victory, nor does it reliably reflect overall performance when considered in isolation. xG is a probabilistic estimate of shot quality, not a holistic measure of team strength, tactical control, or defensive stability. Treating it as such can lead to misleading narratives and inappropriate decision-making [5, 15]. Instead, xG should be used for what it is designed to quantify (shot difficulty and the process of chance creation) providing valuable insight into finishing efficiency, attacking production, and player evaluation in scouting contexts. A more responsible analytical practice requires embedding xG within broader performance frameworks, rather than elevating it to an all-purpose indicator of team merit.

### 6.1.4 RQ4 : How can xG be complemented with additional data-driven indicators to obtain a more complete representation of performance?

The event-based model developed in this work captures a different dimension of team performance, aggregating attacking and defensive behaviours over the full match. Its superior ability to predict match outcomes compared with the xG-SpatialAware model indicates that broader match dynamics (ball progression, pressure patterns, recoveries, and territorial control) play a decisive role that shot-based measures alone cannot reflect. The fact that the event-based model achieves higher predictive accuracy suggests that xG should not be interpreted as a comprehensive measure of performance, but rather as a partial descriptor focused on chance quality.

Importantly, the combined model integrating team-level xG features into the event-based framework yields the highest overall effectiveness. This demonstrates that xG and event-level behaviours encode complementary rather than redundant information: xG quantifies the quality of scoring opportunities, while event-based descriptors characterise the processes that generate and prevent those opportunities. Their integration provides a more complete representation of team strength and offers a principled way to reconcile the long-standing tension between evaluating chance quality and evaluating match dominance.

## 6.2. Limitations

Several limitations of the present study must be acknowledged. First, the use of open-data freeze-frame snapshots restricts the richness of the spatial information available; continuous tracking data would enable more precise modelling of defender and goalkeeper dynamics [2, 15]. Second, the models were trained on a specific set of elite men's competitions, and further evaluation is needed to assess generalization between leagues, playing styles, and women's football [16].

A further limitation concerns the absolute level of predictive accuracy achieved by the proposed models. Although the xG, event-based, and combined approaches attain results that are competitive with (or exceed) those typically reported in the football analytics literature [2, 4, 6, 15], their overall accuracy remains moderate in absolute terms. This is not merely a modelling shortcoming, but reflects intrinsic properties of the problem itself.

Football match outcomes and shot conversions are strongly influenced by human behaviour, contextual contingencies, and stochastic factors that cannot be fully captured by event or positional data alone [3]. Even highly expressive models operating on rich data representations are therefore constrained by the fundamentally noisy and low-scoring nature of the sport. From this perspective, the moderate effectiveness of the models should not be interpreted as a failure to capture meaningful structure, but rather as evidence of the inherent uncertainty embedded in football processes. Importantly, the value of these models lies not in precise outcome prediction, but in their ability to characterize probabilistic tendencies, relative advantages, and systematic patterns of play. As such, the results reinforce the view that football analytics should prioritise calibrated probabilistic assessment and comparative evaluation over deterministic prediction.

## 6.3. Practical applications

The findings of these analyses have direct applications for football analysis. Each model presented in this study is best suited to specific professional use-cases, and its application is conceptually preferable when aligned with the level of performance it is designed to capture.

### 6.3.1 XG-StaialAware model use-cases

First, this study cautions against the use of aggregate xG values as stand-alone indicators of match dominance or team superiority. While xG remains valuable for evaluating shot quality, finishing efficiency, and player-level performance, its interpretation should be contextualised within broader representations of match dynamics, even when the metric effectiveness is at the highest achievable level, like in the case of the xG-SpatialAware model.

In practice, the xG model is best used for shot-level and long-horizon questions: assessing the quality of chances created and conceded, monitoring finishing and goalkeeping performance over larger samples, and supporting player evaluation and recruitment by separating shot volume/quality from conversion variance.

### 6.3.2 Event-based model use-cases

Second, the event-based performance model highlights how match outcomes are strongly influenced by in-game events that are not captured by shot quality alone. Features related to ball progression, pressure intensity, recoveries, territorial control, and possession dynamics provide a more accurate descrip-

tion of teams' performance over the course of a match. Including these dimensions in a game-analysis, analysts can distinguish between matches in which a team creates high-quality chances sporadically and matches in which there was a frequent match-control through pressing, circulation, and territorial dominance. For example, two teams may record similar aggregate xG values, yet exhibit very different performance profiles: one relying on a small number of high-quality chances, the other generating territorial pressure, frequent entries into the final third, and sustained possession. The event-based features considered in this study make these differences explicit, allowing performance to be evaluated with greater clarity than xG alone.
The event-based model is most useful for match-level performance reviews in professional workflows: post-match debriefs, opposition scouting, and preparation tasks where staff need to understand how a team achieved (or failed to achieve) control, which phases of play drove dominance or suffering, and whether the match process was stable even when the shot count is low.

### 6.3.3 Combined model use-cases

Finally, the combined modelling approach offers a principled way to integrate shot-based and event-based information, supporting more precise decision making in performance evaluation, opposition scouting, and post-match analysis. By embedding xG within a richer behavioural framework, practitioners can move beyond outcome-driven narratives and toward a more process-oriented assessment of team performance.
From an applied standpoint, the combined model is the most appropriate tool when a single summary is needed: it retains the interpretability of shot quality through xG while accounting for broader match behaviours, providing an integrated view for reporting, comparative benchmarking across matches, and decision support where both chance quality and overall match control are relevant.

## 6.4. Final conclusions

This study, in the first place, relying new features based on spatial and freeze-frame data, builds a new xG-SpatialAware model able to reach a top-tier level of effectiveness showing how working on certain features and making certain design choices can improve xG metric in absolute terms.
Beyond empirical performance, this work contributes methodologically by providing a transparent and reproducible operational framework for xG modelling. By explicitly defining data preparation, feature engineering, spatial context extraction, validation strategy, and evaluation metrics, the proposed pipeline clarifies how xG values emerge from specific modelling choices and helps reduce ambiguity in their interpretation. While not intended as a universal standard, this framework offers a concrete reference for more consistent and comparable xG implementations, allowing anyone to consult the design choices made and, at the same time, make the metric itself more interpretable, clarifying all the doubts that existed around the metric and how it was constructed.
This study revisits the role of xG within modern football world by analysing its modelling limits and its relationship with broader match dynamics. The results of the xG-SpatialAware model show that even highly engineered shot-based estimators remain inherently constrained by the random nature of football actions. These limitations persist regardless of modelling choices and highlight that xG, while informative, cannot serve as a comprehensive indicator of team performance.
To address this limitation, an event-based performance model was introduced to capture attacking and defensive behaviours accumulated over the full match. The better accuracy of this model in predicting match outcomes demonstrates that match dynamics such as territorial control, pressure patterns, ball progression, and defensive activity play a decisive role beyond shot quality alone.
Additionally, the integration of team-level xG features into the event-based framework yielded the strongest predictive results, confirming that xG and event-level information encode complementary aspects of performance rather than redundant signals.
Empirically, all three models achieve strong predictive effectiveness within their respective scopes: the xG-SpatialAware model reaches top-range shot-level results compared to recent public-data benchmarks, the event-based model provides a competitive match-level signal, and the combined approach yields the most robust overall accuracy. Together, these results support the practical value of the

proposed framework as both a high-performing modelling pipeline and a coherent tool for performance analysis.
Finally, despite the excellent effectiveness of the model, the findings reinforce the view that xG should be interpreted as a probabilistic descriptor of chance quality rather than a deterministic measure of match dominance. When embedded within broader performance representations, xG remains a valuable component of football analytics; when used in isolation, it risks oversimplifying complex match processes. The results encourage a shift toward integrated, process-oriented evaluation frameworks that combine shot-based and event-based approaches to more accurately reflect the already complex random nature of football.

## References


[1] Lightgbm documentation: Advanced topics (missing value handle; categorical feature support), 2025.

[2] Gabriel Anzer and Pascal Bauer. A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living*, 3, March 2021.

[3] Marc Brechot and Raphael Flepp. Dealing with randomness in match outcomes: How to rethink performance evaluation in european club football using expected goals. *Journal of Sports Economics*, 21(4):335–362, January 2020.

[4] Mustafa Cavus and Przemyslaw Biecek. Explainable expected goal models for performance analysis in football analytics. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, page 1–9. IEEE, October 2022.

[5] Jesse Davis and Pieter Robberechts. Biases in expected goals models confound finishing ability, 2024.

[6] H Eggels, R van Elk, and M Pechenizkiy. Expected goals in soccer: Explaining match results using predictive analytics. In *The machine learning and data mining for sports analytics workshop*, volume 16. Technische Universiteit Eindhoven, 2016.

[7] Piotr Florek and Adam Zagdański. Benchmarking state-of-the-art gradient boosting algorithms for classification. *arXiv preprint arXiv:2305.17094*, 2023.

[8] Leander Forcher, Leon Forcher, Alexander Woll, and Stefan Altmann. Ai in bundesliga match analysis—expected possession value (epv) vs. expected goals (xg) to predict match outcomes in soccer. *Frontiers in Sports and Active Living*, 7, November 2025.

[9] Jerome Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 11 2000.

[10] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[11] Stephanie A. Kovalchik. Player tracking data in sports. *Annual Review of Statistics and Its Application*, 10(1):677–697, March 2023.

[12] Lorenzo Lolli, Pascal Bauer, Callum Irving, Daniele Bonanno, Oliver Höner, Warren Gregson, and Valter Di Salvo. Data analytics in the football industry: a survey investigating operational frameworks and practices in professional clubs and national federations from around the world. *Science and Medicine in Football*, 9(2):189–198, May 2024.

[13] Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr, and Iain Matthews. quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. 2015.

[14] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[15] James Mead, Anthony O'Hare, and Paul McMenemy. Expected goals in football: Improving model performance and demonstrating value. *PLOS ONE*, 18(4):e0282295, April 2023.

[16] Sachin Narayanan and N. David Pifer. An xg of their own: Using expected goals to explore the analytical shortcomings of misapplied gender schemas in football. *Journal of Sport Management*, 38(2):92–109, March 2024.

[17] Jonathan Pipping, Tianshu Feng, and R. Paul Sabin. Beyond expected goals: A probabilistic framework for shot occurrences in soccer, 2025.

[18] Al Rahrooh, Anders O. Garlid, Kelly Bartlett, Warren Coons, Panayiotis Petousis, William Hsu, and Alex A.T. Bui. Towards a framework for interoperability and reproducibility of predictive models. *Journal of Biomedical Informatics*, 149:104551, January 2024.

[19] Alex Rathke. An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, 12(Proc2), 2017.

[20] Alexander Scholtes and Oktay Karakuş. Bayes-xg: player and position correction on expected goals (xg) using bayesian hierarchical approach. *Frontiers in Sports and Active Living*, 6, June 2024.

[21] StatsBomb. Statsbomb open data. `https://github.com/statsbomb/open-data/tree/master`, 2025. Accessed: 2025-08-01.

[22] Tuomas Tiippana. How accurately does the expected goals model reflect goalscoring and success in football? 2020.

[23] Dinesh Vatvani. Upgrading expected goals. `https://blogarchive.statsbomb.com/articles/soccer/upgrading-expected-goals/`, 2022. StatsBomb Blog Archive, accessed 2025-12-21.

## Abstract in lingua italiana

I modelli di Expected Goals (xG) sono diventati uno strumento centrale nella moderna analisi calcistica, tuttavia il loro focus esclusivo sulla qualità del tiro impone limitazioni strutturali quando vengono utilizzati per valutare la performance di squadra o prevedere gli esiti delle partite. Questo studio conduce un'analisi in tre fasi, finalizzata a rivalutare criticamente il ruolo e l'affidabilità degli xG utilizzando dati pubblicamente disponibili.
Un primo modello xG ad alte prestazioni è stato sviluppato ampliando l'insieme di feature ingegnerizzate, sfruttando il contesto posizionale dei freeze-frame e descrittori contestuali, per ottenere prestazioni di livello top utilizzando dati pubblicamente disponibili. Una volta assicurato che il modello ottenesse prestazioni migliori rispetto ai modelli attuali, il framework operativo è stato reso pubblico per facilitare l'interpretabilità del modello e le sue limitazioni sono quindi state analizzate e discusse. Per andare oltre queste limitazioni, è stata poi introdotta una metrica di performance complementare basata sugli eventi, che aggrega i comportamenti offensivi e difensivi sull'intera partita e cattura dimensioni della prestazione che vanno oltre l'informazione basata sui tiri. Infine, è stato valutato un approccio combinato integrando feature xG a livello di squadra nel modello basato sugli eventi.
Utilizzando dati storici delle competizioni maschili d'élite, è stato mostrato che il modello basato sugli eventi supera leggermente gli xG nella previsione degli esiti delle partite, mentre il modello combinato raggiunge la migliore performance complessiva. Questi risultati indicano che gli xG da soli non possono spiegare completamente le dinamiche di una partita, ma rimangono altamente informativi quando inseriti all'interno di rappresentazioni più ampie del comportamento di squadra. I risultati evidenziano la necessità di metriche di performance più complete e chiariscono i ruoli complementari delle informazioni basate sui tiri e sugli eventi nell'analisi calcistica.

**Parole chiave:** Expected Goals (xG), Analisi calcistica, Dati evento, Variabili spaziali

## Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Fabrizio Pittorino, for giving me the opportunity to develop a thesis in a field that has fascinated me since childhood. His trust and continuous help allowed me to turn a long-standing passion into an academic research project.
I am deeply grateful to my co-supervisor, Alessandro Falcetta, for guiding me through every step of this work. His constant support and availability made this experience both insightful and productive, and always reassuring.
I would also like to thank the other co-advisors, Alessio Rossi and Carlo Simonelli, for providing valuable perspectives beyond the academic environment. They mentored me throughout this process, shared invaluable insights, and made me even more passionate about this field.
I am thankful to all the people I met during my university years and, in particular, to the friends who have remained by my side and have become part of my life. Sharing this path with them truly made it lighter and pushed me to do better every day.
Finally, I would like to thank my family for giving me the extraordinary opportunity to pursue this path and study what I love most, and above all for supporting me unconditionally through every high and low. I am also grateful to them for instilling in me the desire to keep learning through study and education, and for teaching me the value of responsibility and commitment.
And a special thank you to my girlfriend, for giving me the strength to face every challenge and difficult moment, and for encouraging me to believe in myself a little more every day.
Luca