POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Luca Romano

Advisor: Prof. Fabrizio Pittorino

Co-advisor: Alessandro Falcetta, Alessio Rossi, Carlo Simonelli

Academic year: 2024-2025

1. Introduction

Recent developments in football analytics have been enabled by the increasing availability of event-based and positional data, which has supported a broader shift toward quantitative, data-driven performance evaluation in professional clubs, media, and independent analysis. Within this context, expected goal (xG) metric is framed as a probabilistic estimate of shot conversion: for each attempt, an xG model maps contextual descriptors to a probability between 0 and 1, interpretable through historical conversion frequencies for comparable shots [4]. The thesis was motivated by four research questions: improving xG using publicly available data (RQ1), defining a public operational framework to standardise xG construction and evaluation (RQ2), examining xG's intrinsic statistical and conceptual limitations and its frequent misinterpretations in practice (RQ3), and complementing xG with additional indicators to obtain a more complete match-level representation of performance (RQ4).

To pursue these objectives, the work developed an xG model (RQ1) based on public StatsBomb data [5], integrating freeze-frame positional information and engineered spatial context together with standard contextual descriptors, with the aim of achieving accuracy consistent with the state-of-the-art while supporting more reliable applied use of the metric. A second focus (RQ2) concerns methodological heterogeneity: existing xG pipelines vary in feature sets, preprocessing rules, modelling decisions, and evaluation criteria, producing variability in estimated values and limiting comparability across studies. In response, the thesis made the full xG modelling workflow explicit and reproducible, from data preparation to evaluation, providing a public reference for how xG values emerge from specific modelling choices. The work then motivated a critical analysis of xG interpretation (RQ3), emphasising that xG is often treated deterministically rather than probabilistically, and that match-level narratives derived from shot-level aggregation can be misleading, particularly under small samples and high variance [2]. Finally, the thesis introduced a complementary event-based model and a combined approach (RQ4) integrating team-level xG information with match-level event signals [3] to enrich match-level performance analysis, where xG alone is insufficient.

2. Related literature

The literature reviewed in the thesis positions xG as a probabilistic model of shot success based on contextual variables and traces its evolution from early positional formulations to modern shot-based pipelines [4]. A central direction of development concerns richer representations of shooting context: incorporating information on defender and goalkeeper locations at shot time enables features that capture pressure, obstruction, and the effective shooting window, yielding improvements over location-only models and supporting more faithful estimates of shot difficulty [1]. Within these approaches, work on probabilistic corrections has aimed to improve robustness and interpretability by accounting for player and position effects.

Alongside accuracy-focused advances, the literature consistently highlights limits in how xG is interpreted and used. xG is a probabilistic estimate rather than a deterministic statement about performance, and conclusions are sensitive to sample size because match-level shot counts are typically small. Aggregating xG across attempts yields an expected goals total, but this does not directly map to outcome probabilities and can support misleading match-level narratives when treated as a stand-alone indicator [2]. More fundamentally, xG is defined on observed shots and therefore cannot represent performance dimensions that affect outcomes without necessarily manifesting as shot quality, with evidence indicating that incorporating additional match information and temporal structure improves match outcome prediction beyond shot-based summaries [3].

3. Background

Football analytics mainly uses event data and positional (tracking) data. Event data encode matches as time-ordered discrete actions (e.g., passes and shots) with metadata and pitch coordinates, providing a compact basis for modelling. Tracking data provide time-continuous player and ball locations and describe spatial context more richly, but are harder to process and often less accessible. Freeze-frame snapshots sit between the two by recording all player locations at the instant of a key event, especially at shot time, preserving the local configuration needed to approximate pressure, obstruction, and the effective shooting window.

Expected goals models shot conversion probabilistically: for a shot described by features x, an xG model outputs $\mathrm{xG}(x) = P(y = 1 \mid x)$, interpreted as the long-run scoring frequency for similar shots rather than a deterministic statement about one attempt. Summing shot probabilities yields team-level expected goals, an expected value whose reliability decreases when based on few shots.

Probabilistic accuracy is assessed with log-loss and the Brier score: log-loss penalises assigning high probability to incorrect outcomes, while the Brier score measures the mean squared error between predicted probabilities and the observed binary result. Ranking ability is evaluated with ROC–AUC and PR–AUC, which quantify how well the model orders goals above non-goals across thresholds, with PR–AUC being more informative under strong class imbalance. At match level, outcomes are evaluated as a three-class task using accuracy (fraction of correct predictions), macro F1 (average F1 across classes, weighting them equally), and confusion matrices (counts of true against predicted classes to highlight systematic errors).

4. Proposed solution

4.1. Dataset

The study was built on the publicly available StatsBomb Open Data release, which provides event-level annotations for professional matches together with freeze-frame snapshots at shot time [5]. The dataset spans multiple seasons and competitions and follows a standard event taxonomy including timestamps, event types, player identifiers, body-part information, and spatial coordinates on a 120×80 pitch. The working sample used for modelling comprises 74,915 shots from 2,960 matches.

Two complementary sources of information was exploited. The event stream encodes the full sequence of on-ball actions (passes, carries, pressures, duels, defensive actions, and shots) and provides a structured description of match dynamics. Freeze-frames capture the spatial arrangement of all players and the ball at the instant of each shot, including role information and an indicator of whether a player is a teammate or an opponent. This representation, widely used in contextual shot models, supports engineered variables that approximate defensive pressure, occlusion, and the effective shooting window [1]. All available shot events from men's competitions were included, without manual filtering by shot type or context, and each shot was linked to its associated freeze-frame for spatial feature computation.

4.2. Feature engineering

To approximate contemporary public-data xG accuracy, a feature set was built combining original event descriptors with engineered spatial and freeze-frame variables. Event features capture shot execution and context (play pattern, position, technique/body part, pressure and shot flags). Spatial features describe geometric difficulty and key-pass properties. Freeze-frame features quantify local defensive and goalkeeper context (defender proximity/counts, pressure, goalkeeper positioning) and obstruction-aware measures (visible goal angle/ratio), plus shooter direction. Match-state variables add home/away and pre-shot score context, assisted-shot indicators, and shot-type clusters.
Two prediction tasks are considered: shot-level goal vs no-goal (RQ1–RQ3) and match-level outcome (home win/draw/away win) using representations built from aggregated shot probabilities and event-based variables (RQ4).

4.3. xG modelling framework

Shot conversion was modelled through a gradient-boosted decision tree classifier implemented with LightGBM, selected for its effectiveness on structured tabular data, its capacity to capture non-linear interactions, and its support for heterogeneous feature types, including categorical variables. The resulting estimator was referred to as the xG-SpatialAware model, reflecting its reliance on engineered spatial features derived from freeze-frame context. Model development followed a stratified five-fold protocol at the shot level, with three folds used for training, one for validation-based hyperparameter selection via successive-halving grid search, and one held out for testing. Evaluation was conducted at shot level using probabilistic metrics (log-loss, Brier score, ROC-AUC) and threshold-based metrics, and included comparisons against the StatsBomb xG values and two non-informative baselines under identical splits. Model interpretability was supported through feature importance analysis and SHAP inspection. A match-level baseline was also defined by aggregating team xG values per match and fitting a logistic regression to predict the three-class outcome.

4.4. Event-based match outcome model

To complement shot-based evaluation, an event-based match outcome model was introduced to represent broader match dynamics. The target was the final result in three classes from the home team perspective, and inputs were aggregated descriptors computed separately for home and away teams, covering passing activity, progression, shooting volume, defensive actions, recoveries, pressures, possession dynamics, and related indicators. This model was implemented as a LightGBM multiclass classifier tuned via cross-validation, without using shot-level xG at any stage. Performance was assessed on held-out matches using accuracy, macro F1-score, per-class precision and recall, and confusion matrices, and the model was interpreted as a complementary perspective that reflects cumulative match behaviour rather than shot conversion probabilities.

4.5. Combined model

A combined model was then constructed by adding team-level xG features produced by the xG-SpatialAware estimator to the event-based match representation. This integration was designed to capture both the overall match event profile and the quality of the created scoring opportunities. The combined approach was evaluated with the same match-level protocol and metrics.
Finally, the three approaches were compared in terms of match outcome classification performance and in terms of team–match points prediction. Predicted outcomes were mapped to expected points using the standard 3–1–0 rule, accumulated across matches, and compared against realised totals at team level. Performance was quantified with MAE and RMSE both globally and by competition, providing a longer-horizon assessment of predictive consis-

tency across different competitive contexts.

5. Experimental results

5.1. xG-SpatialAware model effectiveness

The xG-SpatialAware model achieved stable shot-level effectiveness across the five stratified folds, with average log-loss 0.2624, Brier score 0.0751, and ROC–AUC 0.8292, indicating that a substantial share of the available signal in shot outcomes was captured despite the intrinsic randomness of goal scoring. Model explanations based on SHAP confirmed that geometric descriptors dominated the prediction mechanism: base shooting angle and visible goal angle were among the most influential variables, with shot location contributing consistently in the expected direction. Freeze-frame variables related to defensive pressure and goalkeeper positioning provided additional directional contributions, while contextual and match-state features had a comparatively minor role, supporting the interpretation of xG as primarily encoding instantaneous shot conditions rather than broader match dynamics. The results are reported in Tables 1, 2, 3.

Table 1: **Comparison of predictive accuracy between the proposed xG-SpatialAware model, the publicly available StatsBomb xG values [5], two dummy baselines, and representative performance ranges reported in recent literature [1, 2].** Results are reported as mean ± standard deviation over five stratified folds.

Model	ROC–AUC	PR–AUC
xG-SpatialAware (ours)	0.8308 ± 0.0059	0.4889 ± 0.0086
StatsBomb public xG	0.8277 ± 0.0067	0.4784 ± 0.0049
Dummy (most-frequent)	0.5000 ± 0.0000	0.1106 ± 0.0000
Dummy (stratified)	0.5031 ± 0.0024	0.1113 ± 0.0005
Literature (top range)	0.80–0.84	0.32–0.43

Table 2: Comparison of predictive accuracy (continued). Results are reported as mean ± standard deviation over five stratified folds.

Model	LogLoss	Brier
xG-SpatialAware (ours)	0.2616 ± 0.0030	0.0750 ± 0.0007
StatsBomb public xG	0.2642 ± 0.0027	0.0754 ± 0.0005
Dummy (most-frequent)	3.9871 ± 0.0013	0.1106 ± 0.0000
Dummy (stratified)	7.0009 ± 0.0336	0.1942 ± 0.0009
Literature (top range)	0.260–0.270	0.073–0.082

Table 3: Comparison of predictive accuracy (continued). Precision, recall, and F1-score refer to the positive class (goals).

Model	Precision (Goal)	Recall (Goal)	F1 (Goal)
xG-SpatialAware (ours)	0.6970 ± 0.0055	0.2520 ± 0.0090	0.3700 ± 0.0097
StatsBomb public xG	0.7121 ± 0.0161	0.2342 ± 0.0068	0.3525 ± 0.0092
Dummy (most-frequent)	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
Dummy (stratified)	0.1162 ± 0.0043	0.1144 ± 0.0042	0.1153 ± 0.0042
Literature (top range)	0.30–0.74	0.24–0.67	0.35–0.44

When evaluated on identical splits, the proposed model slightly improved over the publicly available StatsBomb xG baseline across all reported probabilistic metrics, while remaining within the upper performance range reported in recent literature [1, 2, 5]. The goal-class precision, recall, and F1-score remained modest for both approaches, reflecting the difficulty of predicting rare scoring events in open data. A complementary experiment highlighted the limited match-level implications of shot-level calibration: the team with higher aggregate xG wins only 61% of matches, reinforcing that xG alone does not reliably translate into outcome-level predictive power.

5.2. Event-based model effectiveness

At match level, the event-based model attained higher three-class outcome prediction accuracy than the xG-only baseline obtained by aggregating xG and fitting a multinomial logistic regression. Across folds, accuracy increased from 0.623 (xG-based baseline) to 0.657 (event-based), with confusion matrices showing clearer gains in identifying home wins and home losses, while draws remain difficult for both approaches. The SHAP analysis of the event-based model indicates that outcome predictions were driven by cumulative match behaviours, with offensive production and territorial indicators supporting the win class, more balanced patterns for draws, and symmetric away-team dominance signals for losses.

5.3. Combined xG-SpatialAware + Event-based model

Integrating team-level xG features into the event-based representation yielded the strongest match-level performance. The combined model reached 0.676 accuracy, improving over the event-based model (0.657) and the xG-based baseline (0.623), and also increased macro F1-score to 0.581, with gains observed across outcome classes and partial improvement in draw recognition. These results indicate that aggregated event descriptors and xG encode complementary, non-redundant information: the event-based model captures broader match dynamics, while xG contributes signal related to the quality of created opportunities. Confusion matrices are reported in Figure 1

Figure 1: **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**

(c) Event-based

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

5.4. Long-term effectiveness and points prediction

A final evaluation considered longer-horizon consistency through expected points prediction. Mapping each predicted match outcome to points under the 3–1–0 rule and comparing predicted versus realised totals yielded lower global errors for the event-based model than for xG alone, and the lowest error for the combined model. Global RMSE decreased from 1.283 (xG-SpatialAware) to 1.169 (event-based) and to 1.103 (combined), with the same ordering reflected by MAE. Competition-level RMSE comparisons show that xG alone is generally the least accurate, the event-based model reduces error across most contexts, and the combined approach further improves or stabilises performance in the majority of competitions, particularly where more matches are available.

The violin plot 2 summarizes the distribution of point-prediction error (RMSE) across random seeds, pooling all competitions with at least 50 matches.

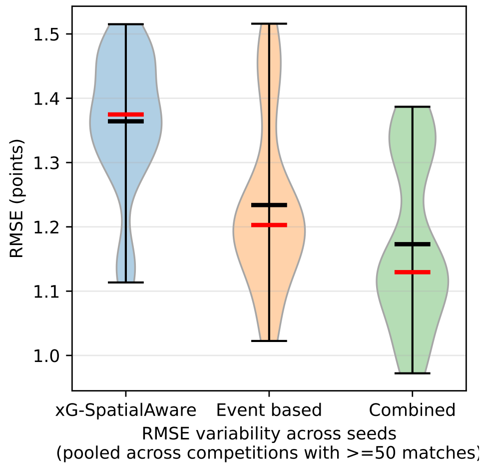

Figure 2: **Violin plots of team–match point-prediction error (RMSE) across random seeds, pooled over competitions with at least 50 matches.** Each violin shows the distribution of RMSE values obtained by repeating model training with different random initialisations for the xG-SpatialAware, Event-based, and Combined approaches. Lower values indicate better long-term estimation effectiveness.

6. Conclusions

The study addresses four objectives aligned with the research questions: improving xG predictive effectiveness with publicly available data, providing a reproducible operational framework, clarifying what xG can and cannot support as an analytical tool, and testing complementary approaches for match-level evaluation where shot-based xG is insufficient.

For RQ1, the results support that xG effectiveness improves when shooting context is enriched with engineered spatial information from freeze-frames. The xG-SpatialAware model combines geometric descriptors and local defensive context, with visible goal features emerging as central drivers alongside shooter position.

For RQ2, the work responds to the lack of an operationally consistent framework in xG modelling by making the pipeline explicit. By specifying data preparation, feature engineering, freeze-frame interpretation, cross-validation strategy, and evaluation choices, it provides a transparent reference workflow, strengthened by the public release of the full implementation.
For RQ3, the analyses show that even highly engineered shot-probability estimators remain constrained by structural limitations, including high variance in finishing and the imperfect representation of spatial interactions under the available data [1]. Aggregate xG is often used as a proxy for match dominance, but higher xG does not reliably imply match victory or a complete description of performance, since xG remains a probabilistic estimate of shot quality rather than a holistic measure of team strength.
For RQ4, the event-based model aggregates attacking and defensive behaviours across the full game and aligns more strongly with match outcomes than the xG-based approach [3]. The combined model integrating team-level xG features achieves the highest overall effectiveness, indicating that xG and event-level descriptors encode complementary information.
Limitations include the restricted spatial richness of freeze-frames compared to continuous tracking data [1], training on elite men's competitions, and moderate absolute predictive accuracy reflecting intrinsic uncertainty not fully captured by event or positional data alone [3]. Accordingly, the models support calibrated probabilistic assessment and comparative evaluation rather than deterministic prediction.
Overall, the thesis develops a top-range public-data xG model through engineered freeze-frame context, provides a transparent and reproducible xG pipeline, and shows that integrating shot-based and event-based perspectives yields the most reliable match-level performance representation, with xG most informative when embedded within broader performance frameworks.

7. Acknowledgements

I would like to thank Professor Fabrizio Pittorino for giving me the opportunity to work on a thesis in a field that has inspired me since childhood, and Alessandro Falcetta for his constant guidance and constructive support throughout the project. My gratitude also goes to Alessio Rossi and Carlo Simonelli for their valuable insights and for introducing me more deeply to the professional side of football analysis.
I am grateful to the friends I met during my university years for making this path lighter and for motivating me to always aim higher. I thank my family and relatives for their unconditional support and for instilling in me the value of education and the desire to keep learning. Finally, I thank my girlfriend for her constant support, for giving me strength in every challenge, and for encouraging me to believe in myself more each day.
Luca

References

[1] 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.

[2] 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.

[3] 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.

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

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