An AI-driven approach to adapting the Expected Goals (xG) model to women's football

Tomasz Lipowski[0009−0002−8085−6628], **Tomasz Piłka**[0000−0003−1206−2076]

Adam Mickiewicz University
Faculty of Mathematics and Computer Science
Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland
tomlip2@st.amu.edu.pl, tomasz.pilka@amu.edu.pl

Abstract. *The Expected Goals (xG) model is a key metric in football analytics, yet conventional models overlook biomechanical and tactical differences in women's football. This study introduces an AI-driven xG model, integrating novel contextual variables like Defensive Congestion Index (DCI) and Shot Block (SB). Using neural networks, the model improves predictive accuracy over traditional approaches. Results show distinct shot conversion patterns in women's football, particularly for long-range attempts, underscoring the need for gender-specific modeling. Incorporating additional features reduces log loss and enhances shot predictions.*
Keywords: *football analytics, artificial intelligence, xG, machine learning models*

1. Introduction

Expected Goals (xG) is a widely accepted metric in soccer analytics that estimates the probability of a shot resulting in a goal, leveraging historical data and contextual factors [1, 2]. Originally a concept from gaming culture, it is now a key tool for teams, analysts, and betting companies to guide performance assessment and strategic planning [3, 4]. By considering elements such as shot distance, angle, and defensive context, xG helps address the low-scoring, unpredictable nature of the sport, providing deeper insights into both team and player performance [5].

Key variables influencing xG include:

- **Shot location**: Distance from goal, angle, and positioning.
- **Shot type**: Footed shots, headers, and set-piece scenarios.

- **Defensive pressure**: Number of defenders near the shooter.
- **Assisting pass type**: Through ball, cross, or cut-back.

Expected Goals transforms match data into an assessment of the probability that a shot will result in a goal, going beyond traditional statistics. It takes into account factors such as shot location, type and defensive pressure to help teams better understand scoring potential and performance.

However, most xG models struggle when applied to women's football as they rely heavily on data from men's matches. Biomechanical differences, such as variations in shot power, trajectory and accuracy, challenge the predictive validity of these metrics. In addition, the slower pace of women's football affects shot-creation dynamics and defensive formations, while tactical variations, including different pressing strategies and defensive setups, further differentiate women's matches from the data on which these models are based.

To address these shortcomings, this research aims to systematically evaluate the need for gender-specific adjustments in xG calculations, identify critical contextual variables that can improve predictive accuracy, and develop an AI-enhanced xG framework that incorporates these variables. By addressing these technical limitations, the study aims to produce a more robust and contextually relevant xG model for women's football.

2. Literature Review

Recent work refines xG models for greater predictive accuracy, for example by incorporating pre-shot event sequences [6] or integrating xG with metrics like xA and xPTS [7]. Advanced machine learning methods further enhance defensive evaluations [8], informing performance analysis, tactics, and recruitment. Bransen and Davis [9] adapt xG models for women's soccer, showing that while men's models can partly transfer, distinct shot patterns in the women's game necessitate tailored approaches.

Crucially, the growing body of research on women's football reveals distinct dynamics that require gender-specific modelling. Studies such [10, 11] have documented the unique physical and tactical aspects of women's matches, from differences in biomechanics and shooting patterns to variations in defensive structures. This evidence supports the call for gender-specific xG models that reflect these differences. The authors also highlight the role of ML in identifying patterns that can prevent injuries, particularly in the context of women's football, where

training loads and physical conditioning play a crucial role in player safety and performance. Using the AI and ML methods highlighted in these studies, football analytics can provide more tailored and accurate insights. These approaches allow analysts to adapt models to the specific needs of men's and women's football, providing a deeper, data-driven understanding of player and team dynamics.

3. Methodology

3.1. Data collection

The dataset for this study was mainly sourced from Hudl StatsBomb[1] Open Data service. StatsBomb offers comprehensive match data, including team and player profiles, possession sequences, individual player actions and event locations, providing a complete perspective of the game's dynamics.

Event data captures every logged action (passes, shots, tackles, dribbles) with precise timestamps and pitch coordinates. Our analysis focused on shots, examining location (distance, angle), pass type, and shooting technique (including foot preference). We also introduced two contextual variables: the ***Defensive Congestion Index (DCI)*** for defenders near the shooter, and ***ShotBlock (SB)*** for opponents in the shot path (see Figure 1).

Attribute	Value
distance (m)	29.88
goalkeeper in shot keeper cone	TRUE
players in shot keeper cone	3
enemy players in shot keeper cone	2
players in the box	6
pressing	0
StatsBomb xG	0.021
my xG	0.012

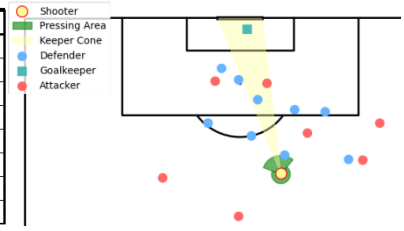

Figure 1: Visual representation of the generated attributes.

For this study, we used StatsBomb Open Data from multiple seasons, encompassing both league and tournament matches. The breakdown of these datasets is presented in Table 1.

[1] https://statsbomb.com/

Table 1: List of leagues and tournaments used for the study.

Competition Name	Year/Season	Gender	Number of Matches	Number of Shots
La Liga	2015/2016	male	380	9071
FA Women's Super League	2018-2021	female	326	8239
FIFA World Cup	2022	male	128	3068
Women's World Cup	2019, 2023	female	116	2891
National Women's Soccer League	2018/2019	female	36	1034
UEFA Women's Euro	2023	female	31	871

3.2. AI-Based Model Development and Evaluation Metrics

The AI-based model development involves two approaches: logistic regression, which serves as the baseline model, and MLP (Multilayer Perceptron) neural networks. An MLP is a type of feedforward artificial neural network that consists of multiple layers of neurons, including an input layer, one or more hidden layers, and an output layer. The MLP model used here consists of three hidden layers with 64, 64, and 32 neurons, followed by an output layer with one neuron using a sigmoid activation function for binary classification. The optimizer used is Adam, and the loss function is binary cross-entropy. Model evaluation is performed using two metrics: log loss, which ensures probability calibration and is fully compatible with logistic regression, and shot outcome comparison, which measures actual versus predicted goal rates.

4. Results and Discussion

Authorial model - *My xG*

As a result of the experiments, it was observed that the use of the same xG model in both men's and women's football leads to **significant errors**, especially for long-range shots, which have higher conversion rates in women's football. The newly introduced attributes show great importance in improving the predictive performance of the model, ranking among the most influential factors, see table 2. Shots, considered as long-distance shots, from the area marked in Fig 2.
Analysis shows that women achieve better results from long-range shots Table 3.

The introduction of **DCI** and **SB** improved the efficiency of the neural network, allowing it to adapt better to different game conditions. The logarithmic loss falls below the level of the StatsBomb model with the original attributes of Table 4, compared to our extended model with additional attributes.

Table 2: Importance of features

Features	**Importance**
distance	0.350
angle degrees	0.307
players in shot keeper cone	0.197
enemy players in shot keeper cone	0.178
players in the box	0.161
pressing	0.156

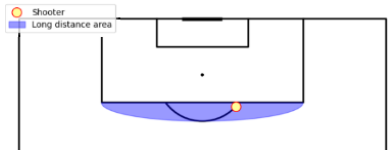

Figure 2: A shot from the blue area or inside the box is not a long-range shot.

Table 3: Results from long-range shots.

	Female	**Male**
Long-range shots	3507	3554
Long-range goals	129	87
Long-range accuracy (%)	3.68	2.45
Long-range frequency (%)	26.90	29.28
Long-range goals/all (%)	0.99	0.72

Table 4: Log loss without and with DCI and SB.

	Original attributes	**Extended attributes**
My xG	0.273	0.262
StatsBomb xG	0.265	0.265

A comparison of the distribution of xG values obtained for the women's and men's football data, respectively, by reconciling the original xG values for the data from StatsBomb and the modification proposed in this paper, are placed in Fig. 3 and Fig. 4, respectively.

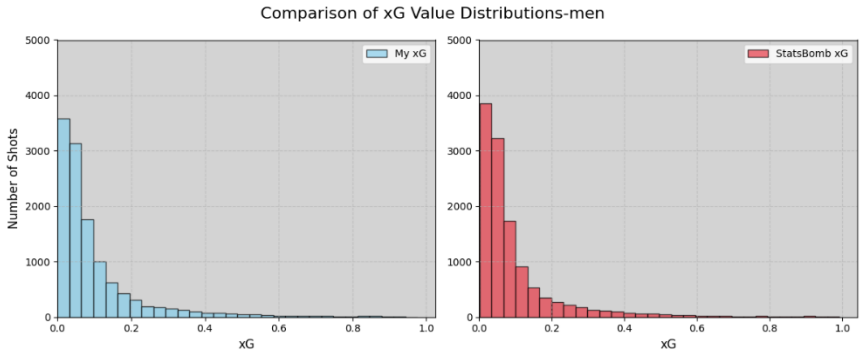

Figure 4: Distribution of my xG proposals vs StatsBomb xG for men.

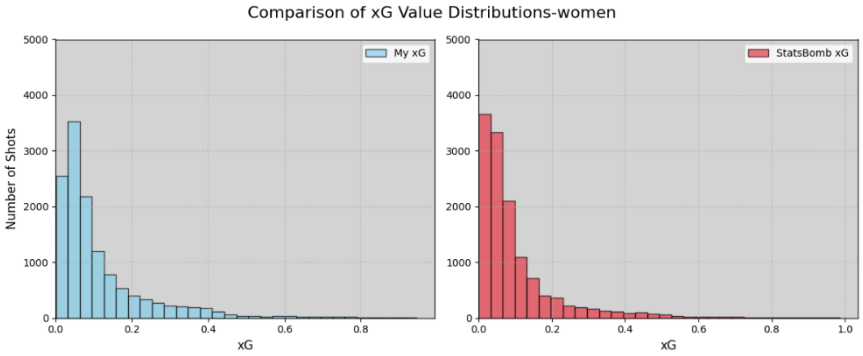

Figure 3: Distribution of my xG proposals vs StatsBomb xG for women.

5. Conclusion and Future Work

This study underscores the need to adapt xG models to the distinct features of women's football. By incorporating AI-driven techniques and contextual variables like Defensive Congestion Index (DCI) and Shot Block (SB), we achieve greater predictive accuracy. Notably, women's long-range shots yield higher conversion rates, highlighting the importance of gender-specific modeling. AI-enhanced methods also address existing xG limitations, providing more reliable tactical insights. Future work should expand datasets with tracking data, integrate computer vision, and develop player-specific models for deeper performance assessments, coaching, and scouting decisions.

References

[1] Statsbomb. What are expected goals (xg)? `https://statsbomb.com/soccer-metrics/expected-goals-xg-explained`, 2024. Accessed: 15-February-2025.

[2] Simpson, M. and Craig, C. Developing a new expected goals metric to quantify performance in a virtual reality soccer goalkeeping app called cleansheet. *Sensors*, 24, 2024. doi:10.3390/s24237527.

[3] Anzer, G. and Bauer, P. A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living*, 3, 2021. doi:10.3389/fspor.2021.624475.

[4] Analyst, O. What Is Expected Goals (xG)?, 2023. URL `https://theanalyst.com/2023/08/what-is-expected-goals-xg`. Accessed: 2025-02-05.

[5] Hewitt, J. H. and Karakuş, O. A machine learning approach for player and position adjusted expected goals in football (soccer). 2023. doi:10.48550/arxiv.2301.13052.

[6] Bandara, I., Shelyag, S., Rajasegarar, S., Dwyer, D., Kim, E., and Angelova, M. Predicting goal probabilities with improved xg models using event sequences in association football. *Plos One*, 19, 2024. doi:10.1371/journal.pone.0312278.

[7] Khrapach, V. and Siryi, O. Statistical metric xg in football and its impact on scoring performance: a review article. *Health Technologies*, 2:47–54, 2024. doi:10.58962/ht.2024.2.3.47-54.

[8] Zaręba, M., Piłka, T., Górecki, T., Grzelak, B., and Dyczkowski, K. Improving the evaluation of defensive player values with advanced machine learning techniques. In *Harnessing Opportunities: Reshaping ISD in the Post-COVID-19 and Generative AI Era (ISD2024 Proceedings)*. 2024. doi:10.62036/ISD.2024.67.

[9] Bransen, L. and Davis, J. Women's football analyzed: interpretable expected goals models for women, 2021.

[10] Pappalardo, L., Rossi, A., Natilli, M., and Cintia, P. Explaining the difference between men's and women's football. *Plos One*, 16:e0255407, 2021. doi:10.1371/journal.pone.0255407.

[11] Eetvelde, H. V., Mendonça, L. D. M., Ley, C., Seil, R., and Tischer, T. Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of Experimental Orthopaedics*, 8, 2021. doi:10.1186/s40634-021-00346-x.