**Abstract:**

This project explores the use of deep learning models to predict Expected Goals (xG) in youth football using player and ball temporal positional data, comparing it to non-temporal models. xG is a metric that estimates the probability of a shot resulting in a goal, based on factors like distance, angle, and pressure from defenders. The research aims to determine whether LSTM models, which capture sequential dependencies, provide more accurate xG predictions than traditional models. Initial results indicate that while the models face challenges with class imbalance, LSTM models show potential for more realistic xG predictions by capturing context and sequential data.

**Intro:**

For many football teams, analyzing player performance is crucial. This allows for more personalized training, identifying weaknesses in a player’s skillset, assessing whether a player is in good form, and much more. However, managing this effectively can be challenging, particularly for clubs with multiple teams, such as a main squad, a B team, and numerous youth teams (e.g., under-23, under-21, under-18, and even multiple teams within the same age category). Ideally, each team would have enough coaches to thoroughly evaluate player performance, but for many clubs outside the top level, this is simply not feasible. One potential solution is to leverage video data by recording matches and training sessions. Coaches or analysts can then review this footage, or advanced AI models can be used to provide detailed summaries. However, this approach is often computationally expensive and cost-prohibitive for smaller clubs. For instance, according to Zone14, the cameras used by Premier League clubs for in-depth video analysis, such as Intel True View systems, can cost up to a million euros(1). This creates a significant financial barrier for smaller clubs or the youth teams of these clubs. A more affordable alternative for performance analysis is the use of positional XY data. This method is both cost-effective and reliable, enabling smaller clubs to gain valuable insights without incurring high expenses. Players can wear GPS trackers, either on their arms or clipped to their shirts, while the ball is equipped with sensors to record critical data such as speed and position. For example, StatSports’ “GPS Performance Tracker” vests, reportedly used by over 800 renowned sporting clubs worldwide (including Liverpool, Arsenal, Manchester City, and the New Zealand All Blacks), are priced at approximately 200 euros per unit(2). These GPS trackers offer several advantages over cameras, such as capturing heart rate, intensity, and providing highly accurate positional data—capabilities that cameras cannot match to the same degree.

One valuable form of analysis in football is assessing a given shot’s Expected Goals (xG). As defined by StatsBomb, a leading sports data company providing analytical services to some of Europe’s top football clubs, xG is “a metric designed to measure the probability of a shot resulting in a goal.(3)”. xG is widely used in performance analysis to assess whether players are over performing or underperforming based on the quality of their scoring opportunities. This is particularly important for youth teams, where players are in development and need to be evaluated to determine whether they are progressing well or require additional attention. xG is applied in various contexts, including: Individual and Team Performance Analysis: Evaluating how efficiently players and teams convert scoring chances. Player Development and Recruitment: Identifying promising talents by analyzing scoring efficiency and consistency. Tactical Insights and Analysis: Understanding shot creation and finishing patterns to optimize strategies. For instance, xG can measure the quality and quantity of scoring chances, offering insights into how many goals a team should have scored based on their opportunities(4). Moreover, it highlights shot quality, helping coaches identify top finishers by revealing which players consistently convert chances at a higher rate than average(5). This makes xG particularly valuable for smaller clubs and youth teams. Many smaller clubs rely on scouting young, underdeveloped, or unscouted players to develop and eventually sell them for significant profit to larger clubs. Similarly, youth players often transition into senior teams, and with better analysis of their development and abilities, clubs can promote outstanding players or identify those unlikely to succeed, ensuring efficient resource allocation. There are multiple methods to calculate xG. As StatsBomb explains, xG models use historical data from thousands of shots with similar characteristics to estimate the likelihood of a goal, represented on a scale from 0 to 1(3). Key factors traditionally incorporated into xG models include: Distance to the goal, Angle to the goal, Body part used for the shot, Type of assist or preceding action (e.g., through ball, cross, set-piece, dribble)(3), as highlighted by Statsbomb. Traditionally, larger clubs use video analysis to calculate xG. However, for smaller clubs, this approach is often too costly and impractical, necessitating more accessible alternatives for leveraging xG insights, like through the use of GPS trackers.

The dataset I am working with is provided by a company called “Forward Football,” which specializes in GPS trackers and data analysis services, primarily catering to smaller clubs and youth teams. While xG is traditionally calculated based on specific shot instances, the actions leading up to the shot can significantly influence its outcome. Factors such as the ball's speed, the trajectory of opponents (e.g., moving toward or away from the ball), whether the player broke through the defense, or if the shot occurred during a counterattack all play a crucial role. These dynamic elements highlight the potential value of incorporating sequential data, such as the ball’s position over time or the player’s movement trajectory, into xG models. This approach could lead to a more accurate and intuitive model by considering the context leading up to the shot rather than focusing solely on the moment the shot is taken. Traditional xG models typically rely on features such as the shot's position, the foot used, or whether it was a free kick. These features can be effectively used in simpler models or machine learning approaches like XGBoost. However, capturing the relationships inherent in temporal and sequential data requires more advanced techniques. Specifically, deep learning models, such as Long Short-Term Memory (LSTM) networks, are well-suited for this purpose.

Given this context, my research question for this project is: “How accurately can a deep learning model predict Expected Goals (xG) using player and ball temporal positional data from youth football games compared to non-temporal data?” My hypothesis is that models utilizing temporal data will predict xG more accurately than those based solely on non-temporal data. This research aims to explore whether the additional context provided by sequential data enhances the predictive power of xG models, particularly in the context

**Related works:**

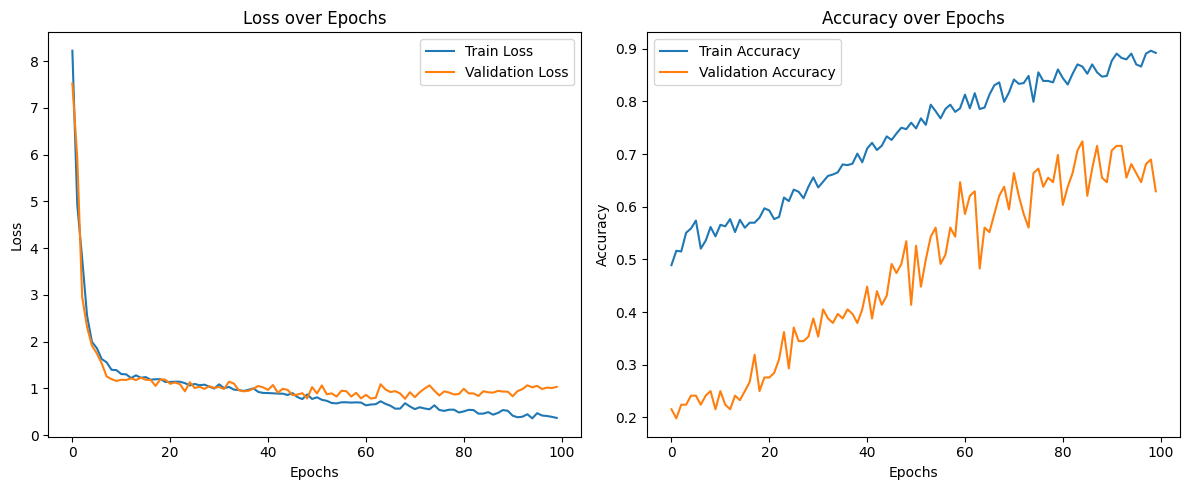
Previous research has demonstrated the success of machine learning in developing models to predict xG values for football players based on their positional data. For example, the study “A Machine Learning Approach for Player and Position Adjusted Expected Goals in Football (Soccer)” by Hewitt et al. (2023) utilized gradient boosting techniques to calculate xG effectively. Additionally, papers such as “The Power of Pixels: Exploring the Potential of CNNs for Expected Goals (xG) in Football” by Matteotti et al. (2024) explored the application of deep learning models for xG calculation, showcasing their effectiveness and the power of convolutional neural networks (CNNs) in this context. The article “Predicting Goal Probabilities with Improved xG Models Using Event Sequences in Association Football” by Ishara Bandara et al. (2024) highlights the importance of incorporating event sequences preceding a shot to significantly improve the accuracy of xG models. This study found that temporal features, such as the "advancement factor" and "player position column," outperformed existing single-event-based models. However, while the research showed promise, it employed a random forest model for feature exploration rather than leveraging the advantages of an LSTM network. The potential of LSTM networks for sequential data tasks is well-documented. For instance, the paper “Classification of Sequential Data in Deep Learning Using LSTM Network” by Padwal et al. (2021) discusses how LSTMs excel at capturing long-term dependencies, making them highly effective for improving classification performance on sequential data. This body of research underscores the potential for combining temporal features with advanced deep learning models, such as LSTMs, to enhance the accuracy and depth of xG predictions in football analysis.

**Methods:**

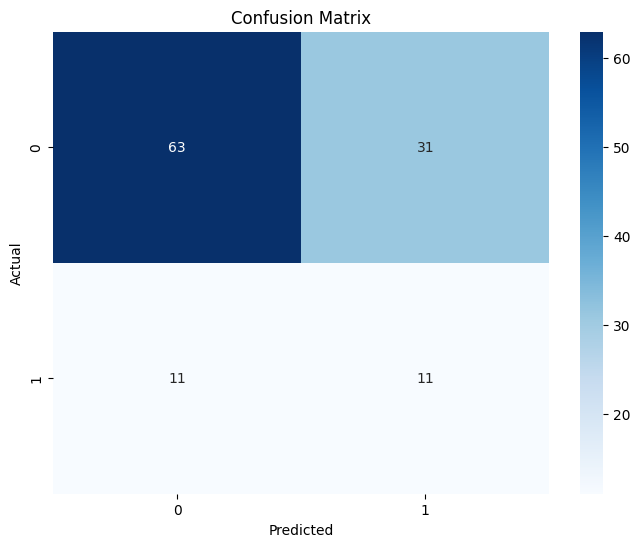
The data for this project comprises two distinct files. The first file records all timestamps when shots were taken during matches, identifies the player who took the shot, and notes whether the shot resulted in a goal. The second file contains 5 Hz positional data for the ball and all players on the field throughout the match. Across all match files, over 600 shot events were identified. To create the sequential data for analysis, I extracted the five seconds of data preceding each shot event. This approach allows the LSTM model to analyze the trajectories of the ball and opponents, uncovering patterns and relationships within the sequential data. From the ball and player **xy-coordinates**, several features were calculated for input into the model. These features were chosen based on StatsBomb's description of key variables in xG models and include: **Distance**: Measures the distance from the shot to the goal. As the distance increases, xG typically decreases. **Angle**: Calculates the angle of the shot relative to the goal. Larger angles generally correspond to lower xG values. **Average Opponent Distance**: Represents the average distance of opponents to the shot taker, simulating the “pressure” the shooter experiences. Higher pressure (i.e., shorter distances) usually reduces xG. **Opponents in Trajectory**: Counts the number of opponents in the direct path of the shot. While the average opponent distance may be high, multiple defenders in the trajectory may block or obstruct the shot, lowering the xG. This methodology departs from traditional xG calculations in a few ways. For example, specific "gamestate" features are not explicitly included. However, the LSTM compensates for this by interpreting sequential data. For instance, the speed and movement of the ball—captured through changes in its position over time—can infer game dynamics like passes or crosses. Similarly, if the ball remains stationary before the shot, it may indicate scenarios like free kicks or corners.

For the baseline model to compare against the temporal model, I used a fully connected layer (FCL) model. Since the data is spatial but not encoded in a way that suits convolutional neural networks (CNNs), an FCL model is more appropriate for achieving reliable results with this dataset. For the temporal model, I incorporated an LSTM layer into the FCL architecture. This addition enables the model to process sequential data and uncover meaningful relationships and patterns over time. The dataset exhibits a significant class imbalance, with approximately five shots for every goal. This imbalance poses a challenge for the model's ability to learn effectively across both classes.To address this, I employed a combination of **SMOTE (Synthetic Minority Oversampling Technique)** and **class weighting**. **SMOTE** addresses class imbalances by generating synthetic samples of the minority class. Its effectiveness, particularly in combination with LSTM models, has been demonstrated in various applications, such as the study by Efendi et al. (2024), “DBSCAN SMOTE LSTM: Effective Strategies for Distributed Denial of Service Detection in Imbalanced Network Environments”. **Class Weighting** assign higher weights to the underrepresented class during training, this approach ensures the model pays more attention to the minority class, reducing the impact of imbalance on model performance. These methods, used together, aim to mitigate the effects of class imbalance, enabling the models to learn more effectively and produce balanced predictions.

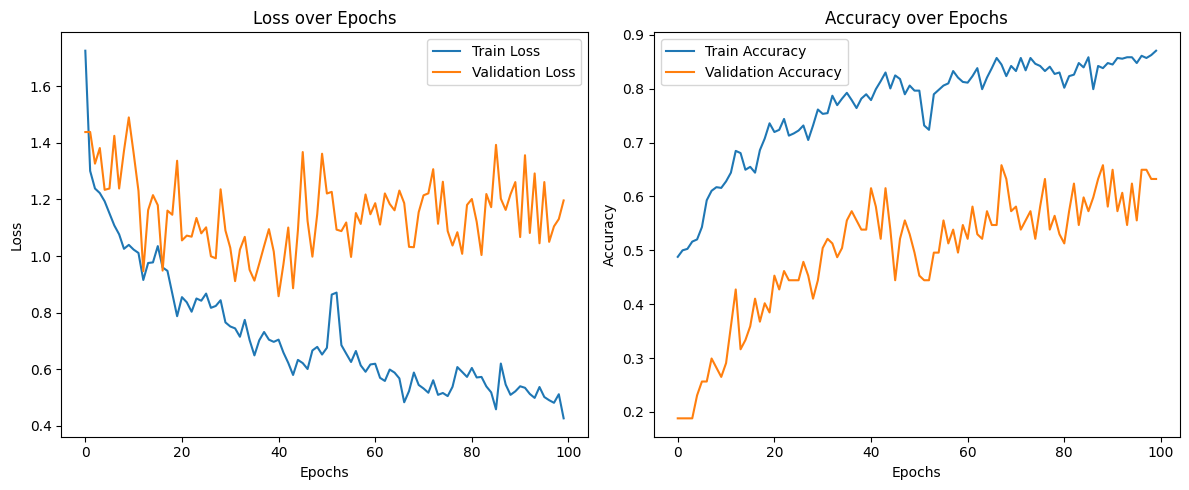
**Results:**



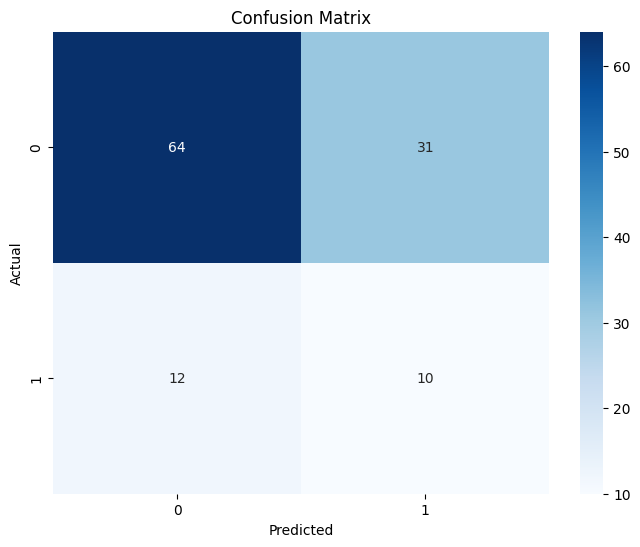
For the baseline FCL model, the training loss shows consistent improvement, indicating that the model is effectively learning from the data. However, the validation loss begins to fluctuate around epoch 60 and even increases significantly during later epochs. Validation accuracy improves steadily up to epoch 75, but starts to exhibit fluctuations and minimal improvements after epoch 60. Earlier epochs, such as epoch 60, also demonstrate good validation accuracy without the subsequent spikes in validation loss. The gap between training and validation performance around epoch 60 is a clear indicator of overfitting. To mitigate this, the limit number of training epochs to 60 during testing to achieve a balance between learning and generalization.



The results after 60 epochs show an accuracy of 0.64, with a macro average of 0.55 and a weighted average of 0.67. The macro average highlights the poor performance of the minority class, while the weighted average, influenced by the dominance of class 0, skews closer to its metrics. These results emphasize the significant impact of the class imbalance in the dataset, as reflected in the confusion matrix. Despite employing techniques to address this issue, the model's performance reflects the persistent challenges posed by the imbalance, indicating the need for further refinement or alternative approaches.

****

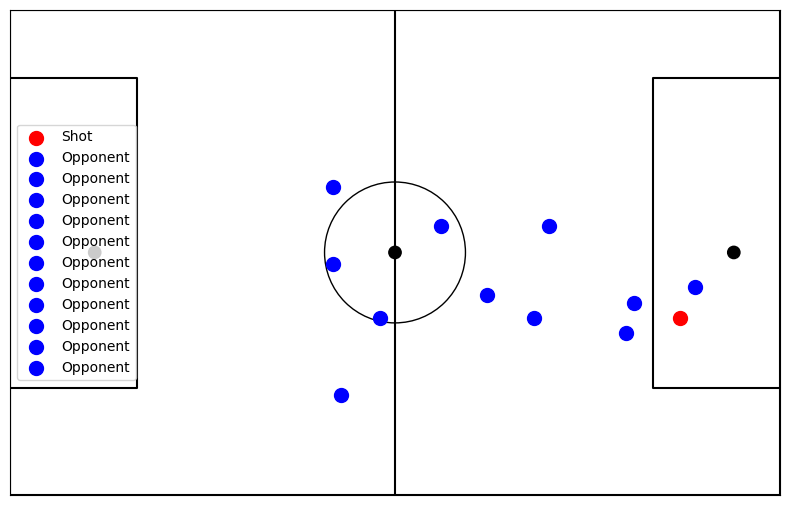
For the LSTM model, validation accuracy improves up to approximately epoch 40, but begins to oscillate without significant improvement afterward. Similarly, validation loss initially decreases but starts fluctuating and increasing beyond epoch 50, strongly indicating overfitting. The best balance between training and validation performance is observed around epochs 40–50, as further training yields no meaningful improvement in validation accuracy and worsens validation loss. These results suggest that the model starts overfitting to the training data beyond epoch 50. Therefore, training will be stopped after 50 epochs to prevent overfitting and achieve optimal performance.



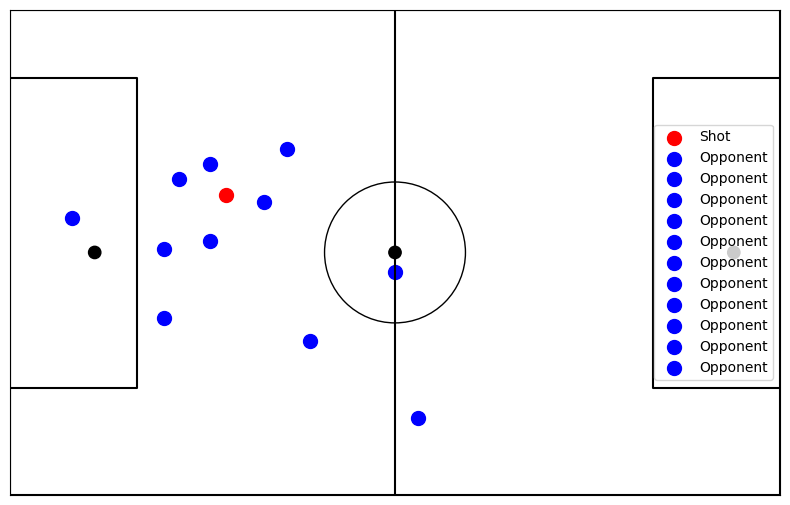
The results after 50 epochs showed an accuracy of 0.63, with a macro average of 0.53 and a weighted average of 0.67. Similar to the previous model, the macro average highlights the poor performance of the minority class, while the dominance of class 0 skews the weighted average closer to its metrics. This indicates that the model struggles with the same issues as the first model, primarily stemming from the significant class imbalance in the dataset. This is further shown in the confusion matrix, which closely resembles that of the FCL model.

**Discussion:**

While the results may initially suggest that the model underperformed due to issues such as persistent class imbalance or a lack of effective training, a closer inspection of the xG predictions reveals a different perspective.



For instance, in Figure 1, the shot was taken close to the goal with minimal pressure and only one player in its trajectory (presumably the goalkeeper). The FCL model correctly predicted the shot as a miss, assigning an xG of 0.23, whereas the LSTM incorrectly predicted it as a goal with a much higher xG of 0.71. Upon inspecting the xG values predicted by both models, it becomes evident that, while the LSTM appears to struggle in traditional accuracy metrics, it generates more realistic xG values. This suggests that the LSTM model's ability to capture sequential dependencies and contextual patterns in the data allows it to better evaluate the quality of scoring opportunities. In contrast, the FCL model, lacking sequential context, may struggle to identify these intricate patterns, limiting its predictive depth.



This trend is also evident in the reverse scenario, where the FCL model correctly predicted the label (goal) while the LSTM did not. In this instance, the shot was taken from a considerable distance—unusual for a typical goal in football—with multiple defenders nearby applying pressure. Such a scenario naturally results in a low xG. This is reflected in the predicted xG values: the FCL model assigned a highly improbable 0.92 (indicating an almost guaranteed goal), whereas the LSTM predicted a much lower and more realistic xG of 0.01. Although the LSTM’s prediction appears very low, it aligns more closely with the expectations for a shot of this nature compared to the overly high FCL prediction.

The class imbalance in the data remains a persistent challenge due to the inherent nature of football, where shots significantly outnumber goals. While this imbalance is unavoidable, collecting more data and employing advanced class balancing techniques could help mitigate its impact and improve model performance.

Although GPS trackers provide precise information on player and ball positions and speeds—critical inputs for deep learning models—they lack certain contextual details that video analysis can offer. Video data adds a third dimension (z-axis), information on which foot was used for the shot, the game state (e.g., corner, free kick), and the angle of the player relative to the ball. Moreover, GPS trackers, often attached to a player's body or arm, may not perfectly reflect the position of the leg or ball, leading to slight inaccuracies in xG calculations. Combining camera footage with GPS data would likely yield the most comprehensive and accurate input for xG models.

Another consideration is the development of specialized xG models tailored to different phases of play, such as free kicks, corners, normal play, and counterattacks. These game states have distinct characteristics and challenges. For example, shots from corners typically occur close to the goal but are harder to convert due to defensive pressure and the use of non-dominant feet. While the sequential nature of LSTM models captures some of this contextual information, dedicated models for each game state would better account for these differences. However, this approach was not feasible for this project due to insufficient data and the absence of game state labels.

**Conclusion:**

In conclusion, this research explored how accurately a deep learning model can predict Expected Goals (xG) using player and ball temporal positional data from youth football games compared to non-temporal data. Both the fully connected layer (FCL) and LSTM models performed similarly in accuracy (around 0.63–0.64), but the LSTM model better captured the dynamics of xG predictions by incorporating sequential data. While class imbalance affected model performance, the LSTM model provided more realistic xG predictions, especially in complex shot scenarios. Further data refinement and balancing techniques are needed to improve accuracy, but the results suggest that temporal data can enhance xG prediction.

**Reference:**

Hewitt, J. H., & Karakuş, O. (2023). A machine learning approach for player and position adjusted expected goals in football (soccer). *Franklin Open*, *4*, 100034.

Matteotti, M., & Sotudeh, H. The Power of Pixels: Exploring the Potential of CNNs for Expected Goals (xG) in Football.

Bandara, I., Shelyag, S., Rajasegarar, S., Dwyer, D., Kim, E. J., & Angelova, M. (2024). Predicting goal probabilities with improved xG models using event sequences in association football. *Plos one*, *19*(10), e0312278.

Padwal, N. D., & Alaskar, K. Classification of Sequential Data in Deep Learning Using LSTM Network.

Efendi, R., Wahyono, T., & Widiasari, I. R. (2024). DBSCAN SMOTE LSTM: Effective Strategies for Distributed Denial of Service Detection in Imbalanced Network Environments. *Big Data and Cognitive Computing*, *8*(9), 118.