STM JOURNALS

International Journal of
Data Structure Studies

ISSN: 3049-2440
Volume 3, Issue 2, 2025
July–December
DOI (Journal): 10.37591/IJDSS

https://journals.stmjournals.com/ijdss

Research IJDSS

Semantics Analysis of Expected Goals in Soccer Data Using Machine Learning

Anuj Razdan*

Abstract

In recent years, the increasing availability of soccer data has greatly enhanced the accuracy and depth of player performance evaluation. Soccer, being one of the most popular sports worldwide, attracts millions of fans due to its simple rules, minimal equipment requirements, and high entertainment value. However, analyzing an entire match manually can be time-consuming, leading to a growing demand for automated methods that can summarize and interpret game data efficiently. This study focuses on extracting, processing, and analyzing soccer data from a CSV file to evaluate player performance and compute expected goals (xG). The approach employs a Gradient Boosting Classifier to assess predictive performance, with the F1 score used as the primary evaluation metric. Additionally, the analysis includes player-specific performance metrics and team-based insights, offering a comprehensive understanding of individual contributions and match outcomes. The integration of machine learning and statistical modeling in this work demonstrates how data-driven techniques can provide valuable insights into soccer analytics, supporting coaches, analysts, and fans in making informed decisions.

Keywords: Soccer data, player, semantics analysis, machine learning, data preprocessing

INTRODUCTION

Soccer is one of the most popular sports in the world due to its simple rules and limited equipment. Watching the entire match takes time and causes many sports fans to enjoy the content. Therefore, the analysis of soccer videos has attracted the attention of extensive research and many applications have been investigated. Analysis of soccer has always been time-consuming, highlighting the need for technology. It is still difficult to evaluate performance in sports due to the wide variety of sports branches. Previous researchers have developed a number of methods for recording highlights in specific sports. For example, some have proposed techniques to detect match and break conditions for content creation, while others have used slow motion and detection tools with SVM classifiers in football videos [1].

Video semantic analysis is complex and growing as football videos become the focus of academic and commercial research. The challenging task of identifying and determining the importance of sports videos is due to their large audience and significant commercial potential. This focus is essential to advances in video retrieval, summarization, personalized recommendation systems, and adaptive content delivery.

This important technology is not only the gateway to unlocking the full potential of football video content, but also the path to new applications that meet the diverse needs of the audience. From increasing the accessibility and relevance of content reach to delivering meaningful and engaging content, the ability to identify and mark important content in football videos is essential for shifting audience.

***Author for Correspondence**
Anuj Razdan
E-mail: kachrurohan1999@gmail.com

Assistant Professor, Department of Computer Science & Engineering, Echelon Institute of Technology, Faridabad, Haryana, India

Received Date: July 08, 2025
Accepted Date: October 08, 2025
Published Date: October 24, 2025

Citation: Anuj Razdan. Semantics Analysis of Expected Goals in Soccer Data Using Machine Learning. International Journal of Data Structure Studies. 2025; 3(2): 31–47p.

As researchers investigate this complex area, they are addressing issues arising from the positive nature of football, where critical moments can escalate rapidly. The intersection of education and business in this pursuit means working together to leverage technologies such as artificial intelligence and machine learning to solve the complexities of football video analysis.

The impact of successful highlight detection and labeling extends beyond mere entertainment, influencing personalized recommendation systems that cater to individual preferences. Moreover, it plays a pivotal role in adaptive content transmission, ensuring seamless and tailored viewing experiences across diverse platforms.

In essence, the quest to master the detection and labeling of soccer video highlights is not merely a technological challenge; it is a gateway to a myriad of possibilities that promise to revolutionize the way they engage with and consume sports content. As researchers continue to innovate and refine these technologies, the landscape of soccer video analysis is poised for transformative advancements, ushering in an era of enriched viewer experiences and unprecedented commercial opportunities.

LITERATURE REVIEW

Various literature reviews about the video are studied thoroughly. The many papers regarding the issue help us understand how to fix it. The numerous papers in this study give us information on the algorithm. It can assist us in using the algorithm to address our current issue.

According to Wang, Semantic analysis is based on fusion of audio/visual features for soccer video [1]. In this study, audio and visual features were combined to improve the semantic analysis of football videos. It uses HTC to extract important content with semantic domain derived from the combination of emotional arousal factors. HFV is designed to break down the main points into goals, shots and fouls. Experimental results show that the proposed specification method is both valid and accurate. The happiness level calculated by HTC can be used to measure content quality. The results of this study can be used for event retrieval and user-centered summarization. Despite this progress, capturing better content from different types of football videos is still a challenge [1]. Our future work will focus on creating better representations to improve event detection and extend this process to other sports videos.

The work of Zhou *et al*. explores audiovisual segmentation (AVS), focusing on creating pixel-level segmentation masks for objects heard in videos. To support AVS research, they developed the Audiovisual Segmentation Benchmark (AVSBench) and extended it to include single-site, multi-site, and text semantic subsets. They investigated three performance domains: semi-supervised single-site AVS (S4), fully supervised multi-site AVS (MS3), and fully supervised audiovisual semantic segmentation (AVSS). As a strong foundation for these sites, they introduced a new pixel-level system that uses the TPAVI module to encode pixel-level audiovisual interactions in video segments and uses constants to enhance listening engagement. Their method demonstrates a strong correlation between audio and visual quality compared to many other methods in AVSBench. Future work includes the development of large-scale electronic devices for pre-training models [2].

As stated by Oskouie *et al*. in "Multimodal Feature Extraction and Fusion for Semantic Mining of Soccer Video: A Survey", published in Artificial Intelligence Review, this study provides an overview of football video research. Current systems are evaluated in the areas of key detection, video conferencing and recovery, ball and player tracking for matches, skill evaluation and use in football analysis [3]. Additionally, various sectors using video analysis tools are also introduced and compared. Many computer vision techniques have been discussed and compared to solve the challenge of providing automatic and real-time video [3].

As stated in "HMM Based Soccer Video Event Detection Using Enhanced Mid-Level Semantic", an effective hidden Markov models (HMMs) based approach for soccer video event detection within a hierarchical video analysis framework is detailed, published in Multimedia Tools and Applications.

Football video footage is divided into four different environments: international, intermediate, close, and viewing. International and regional movement data complete this intermediate point. Normal football videos are divided into event videos, and HMM is used to determine event types by combining time lapses and general features of event clips [4]. Then it compared the significance detection method of Dynamic Bayesian Network (DBN), Conditional Random Field (CRF) [4], and HMM-based method, and the latter achieved better average F value score of 82.92%, and DBN and CRF increased by 9.85 and 11.12% respectively. It also investigated the effects of the number of hidden states, aggregate features, and intermediate semantic refinement on event detection performance [4].

Automated event description is crucial to creating quality video game content. Researchers around the world are looking for effective solutions to detect and classify important events in various sports. Popular methods now use rules based on manual analysis and heuristic knowledge, modeling the same pattern of performance in sports. Machine learning could be another way to bridge the gap between these processes. The combined method proposed includes statistical models with legal methods for discovering values, pioneering the use of competitive games, and methods related to sports such as football, basketball, and Australian rules football. Tests on many sports-equipment have proven the effectiveness and power of this method [5].

A novel framework for semantic annotation and personalized retrieval of sports video is presented by Xu *et al*. Unlike traditional methods that rely heavily on audio or visual features, this model combines web ads with video reviews. This integration increases the accuracy of event detection, clarifies event boundaries, and resolves challenging situations. The approach also enables the creation of personalized content from multiple perspectives related to a specific game, event, player, or team. Frameworks that include text analysis, video analysis, text/video matching, and identification have been shown to be effective in retrieving regions and identities. Overall, this retrieval method effectively helps meet users' expectations [6].

The study by Yu *et al*. introduces a methodology for detecting soccer events by collaboratively analyzing textual, visual, and auditory components. The fundamental concept involves breaking down a match video into smaller segments until the desired eventful segment is pinpointed. The analysis involves the utilization of basic features such as minute-by-minute summaries from sports websites (text), semantic categorization of shots from wide and close-up perspectives (visual), and low-level characteristics such as pitch and log-energy (audio). The study demonstrates that even with the consideration of simplistic features and the avoidance of labeled training data, event identification can be accomplished with remarkable accuracy. Experiments conducted on approximately 30 h of soccer footage exhibit highly promising outcomes in the detection of goals, penalties, yellow cards, and red cards [7].

Sports movies attract the attention of audiences all over the world. Research by Xu *et al*. in this area has focused on exploring what is happening in sports videos to facilitate accessibility and navigation. Most detection methods in sports videos are based on visual features. However, sound, which is the main element of video games, is also important in researching semantic phenomena. In this study, the concept of "content" in paper mining is used to describe different types of data [8]. These special sounds are related to the behavior of players, referees, commentators, and spectators and become the main focus of the famous event. Unlike simple features, audio content can be considered as an intermediate representation that facilitates semantically in-depth analysis. Audio content is derived from simple audio using support vector machine learning. With the help of video images, audio content can be learned by Hidden Markov Model (HMM) to identify events in sports videos. Experimenting with keyword generation and subsequent search results was very helpful. Based on these findings, it is thought that the keyword of sound is a good representation that can provide good results in detecting the phenomenon in sports videos. Actions in three types of guidance demonstrate the feasibility of the proposed method [9].

In the study by Zhang *et al.*, a technology based on self-correction of both threshold and noise is shown to be important for detecting clips in football videos. In this way, a sliding window is used when calculating the main face of the frame in the HSV color space, creating two local, self-correcting fields obtained from the color histogram. It uses the cylindrical distance to extract the primary color pixels and determine the primary color ratio of the frame. It analyzed the shots in football videos by combining two thresholds and color schemes. Experimental results confirmed the effectiveness of this technology in identifying clips and gradually modifying them, thus providing a solid basis for recovering videos [9].

There is an urgent need to devise a strategy for video retrieval based on content, aiming to streamline the management and swift navigation of vast video datasets. Shot categorization stands as a pivotal focus in the processing and retrieval of football videos. Addressing the limitations of existing techniques, it introduces an innovative method capable of categorizing football video shots into primary, intermediate, close-up shots, and others, leveraging sub-region analysis. The crux lies in computing the ratio of field-colored pixels within sub-regions of football video frames in the HSV color space. Experimental outcomes demonstrate the remarkable precision and recall rates of the proposed approach [10].

Playback sequences serve as pivotal markers denoting notable instances in sports reels. Typically, a playback sequence is flanked by twin emblems, serving as beacons for the beginning and conclusion of the sequence. These emblematic transitions, spanning approximately 10–30 frames, delineate the motion of airborne or fluctuating entities. The outlined method comprises two principal phases: initially, an unsupervised acquisition of emblem transition templates occurs, alongside the precise extraction of a pivotal frame (termed K-frame) and a cohort of pixels accurately portraying emblematic features (designated as L-pixels); subsequently, this acquired knowledge is collectively harnessed for the identification of emblems and playback sequences within the visual content. Additionally, optical components are used to identify the movement patterns of the logo along with traditional color analysis. Rigorous experiments demonstrate the effectiveness of the proposed method in identifying logos and performances of various sports [11].

Recent years witness rapid advancements in football video retrieval and abstraction techniques. The extraction of highlights holds significant academic and practical value within the realm of soccer video analysis. Through a novel approach, leveraging Hanjalic's theory of affection curve, a highlights extraction system is proposed, focusing on the emotional fluctuations of the audience. By extracting arousal features from football footage, an arousal model is constructed, facilitating the extraction of highlights. Improvements over Hanjalic's methodology are suggested, notably the incorporation of shot intensity as a pivotal feature for arousal modeling, enhancing the system's recall rate, precision, and computational efficiency. Domain expertise guides the process of locating highlights with precision, further customizable based on user preferences regarding viewing duration. Experimental results confirmed the effectiveness of the plan, affirming the viability of shot intensity as a substitute for motion intensity in arousal modeling, and attesting to the efficacy of the proposed algorithms [12].

According to the research conducted by Hanjalic, modeling addresses the challenge of automatically extracting highlights from TV broadcasts [13]. It seeks a generic method that does not need models for specific events considered highlights by users. Instead, Hanjalic identifies the main market segment that attracts the most users. It is possible to think that the importance of events, unlike other random events, will increase users' satisfaction [13]. It tracks changes in the user's happiness by monitoring the behavior of selecting audio-visual low-level features and video editing schemes over time [14]. The relationship between extra-content and arousal derives in part from psychophysiological research and the practice of live video teaching. The desired change in user happiness is represented by the time happiness curve, given effort preference, modification is made to eliminate overhangs in overall length. Manfredi evaluates and discusses the effectiveness of three methods of using football media [14].

METHODOLOGY

The approach for Semantics analysis of Sports Data Video using XG Model and Gradient Boosting Classifier involves the development of the model that encodes visual, textual audio data simultaneously.

The key steps in the methodology are as follows:

- *Data Preprocessing:* Large multimodal files containing league data descriptions are compiled. Preprocessing is performed on the data to transform text into numerical representations such as word embeddings or tokenized sequences. In this study, it takes the Kaggle Score data [14].
- *Model:* The proposed framework adopts transformer-based architectures to process visual, textual, and auditory inputs simultaneously. Expected Goals Models aim to predict the probability of a particular shot leading to a goal [14]. This metric offers insights into a game beyond just the final score, capturing the dynamics of goal-scoring opportunities. Since goals are the ultimate decider in a match and are typically the outcome of shots, xG metrics primarily rely on shot data [14]. Essentially, any factors affecting a team's "expected goals" tally usually stem from their shot volume [14]. For instance, while it might seem logical that Team B's three red cards would boost Team A's goal-scoring chances, the reality is that having more players on the field leads to more shots, thereby increasing expected goals [14]. Consequently, the Gradient Boosting Classifier, a robust algorithm, focuses solely on shots and their attributes [14]. It employs a collection of decision trees, which, to mitigate overfitting, generates numerous alternative trees using diverse predictors and samples. This strategy helps balance the bias-variance tradeoff, ensuring more stable predictions [14].
- *Contrastive Learning:* Contrastive learning is employed to train the model. It promotes the model to map comparable audio and video characteristics together in the embedding space, while pushing dissimilar pairings apart.
- *Evaluation:* The model's efficiency is evaluated employing conventional retrieval-oriented metrics like precision, recall, and mean average precision. Its proficiency in retrieving precise audio content from videos based on textual queries undergoes thorough scrutiny. Additionally, graphical analysis is employed for the evaluation process.
- *Results:* The results of the experiments show that the suggested semantics analysis of Soccer Data using Machine Learning beats existing approaches. It delivers higher precision and recall rates, making it a useful tool for a wide range of multimedia applications.

The methodology demonstrates the sports data to retrieval of player analysis and expected goals, resulting in a solid solution for analysis of Soccer Data using Machine Learning.

Flow Chart

Figure 1 shows the following process in the flow chart, which is explained below:

- *Soccer Data:* In this study, it takes the Kaggle Score data [14].
- *Data Preprocessing:* Large multimodal files containing league data descriptions are compiled. Preprocessing is performed on the data to transform text into numerical representations such as word embeddings or tokenized sequences. In this study it takes the Kaggle Score data [14].
- *Model:* The proposed framework adopts transformer-based architectures to process visual, textual, and auditory inputs simultaneously. Expected Goals Models aim to predict the probability of a particular shot leading to a goal [14]. This metric offers insights into a game beyond just the final score, capturing the dynamics of goal-scoring opportunities. Since goals are the ultimate decider in a match and are typically the outcome of shots, xG metrics primarily rely on shot data [14]. Essentially, any factors affecting a team's "expected goals" tally usually stem from their shot volume [14]. For instance, while it might seem logical that Team B's three red cards would boost Team A's goal-scoring chances, the reality is that having more players on the field leads to more shots, thereby increasing expected goals [14]. Consequently, the Gradient Boosting Classifier, a robust algorithm, focuses solely on shots and their attributes [14]. It employs a collection of decision trees, which, to mitigate overfitting, generates numerous

alternative trees using diverse predictors and samples. This strategy helps balance the bias-variance tradeoff, ensuring more stable predictions [14].

- *Evaluation:* The model's performance is assessed using typical retrieval-oriented metrics such as precision, recall, and mean average precision. The model's ability to retrieve specific audio content from videos using text queries is evaluated. It also uses graphics analysis for the evaluation.

EXPERIMENTS AND RESULTS

In this Result we used Excepted Goals, we also do player analysis on different goals year wise and country-wise. In this we do player analysis on different scenarios like worse shooter and best finisher for long range. Due to which we are able to understand the wide range of capabilities of our players, it helps to understand in which scenario they played best.

As shown in the Figure 2 given, it is analyzing Shots in football, it will start by looking at some of the game's characteristics. First, let us look at how the various outcomes of a shot are dispersed.

As depicted in Figure 3, adversaries obstruct the majority of attempts; however, this is attributable to the fact that all unblocked projectiles are categorized into numerous groups [14]. Most of the shots that evade blocks are either aimed towards the goal's center or veer off to the left or right side [14]. Now let us focus on an important aspect of the xG model: the percentage of goals scored. It analyzes this data by group and year to identify changes in trends across regions or over time.

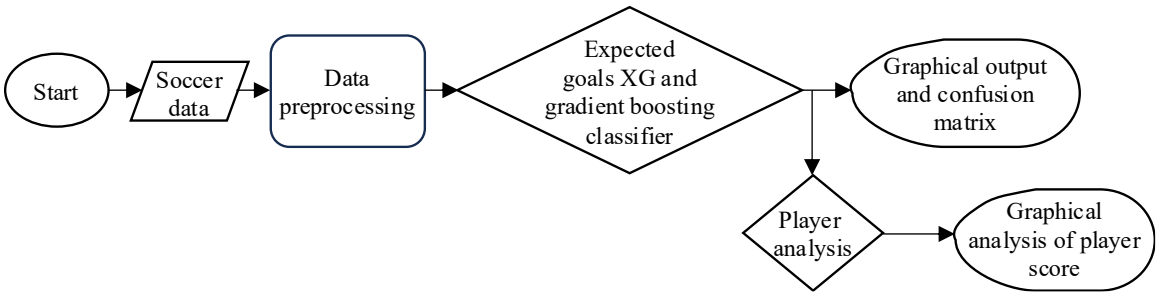

Figure 1. A Flow chart of semantics analysis of soccer's data using machine learning.

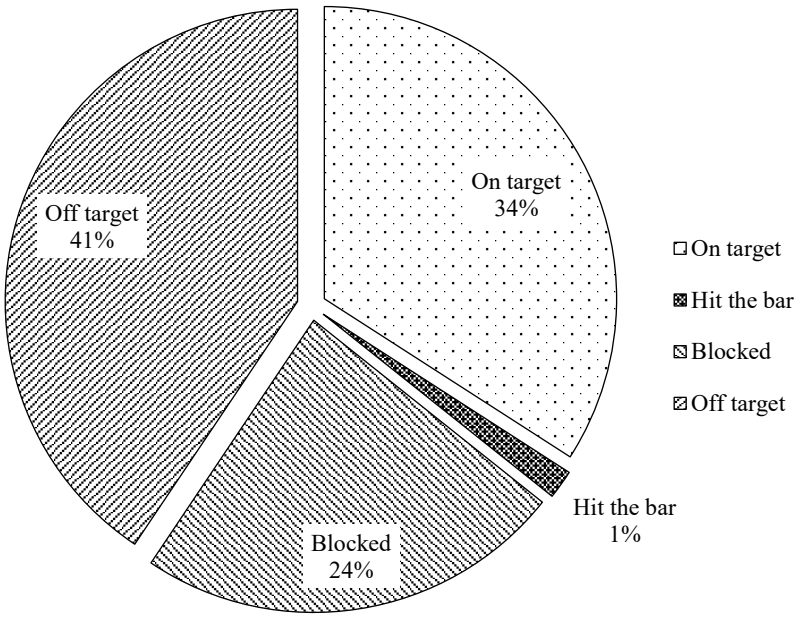

Figure 2. Shot outcomes.

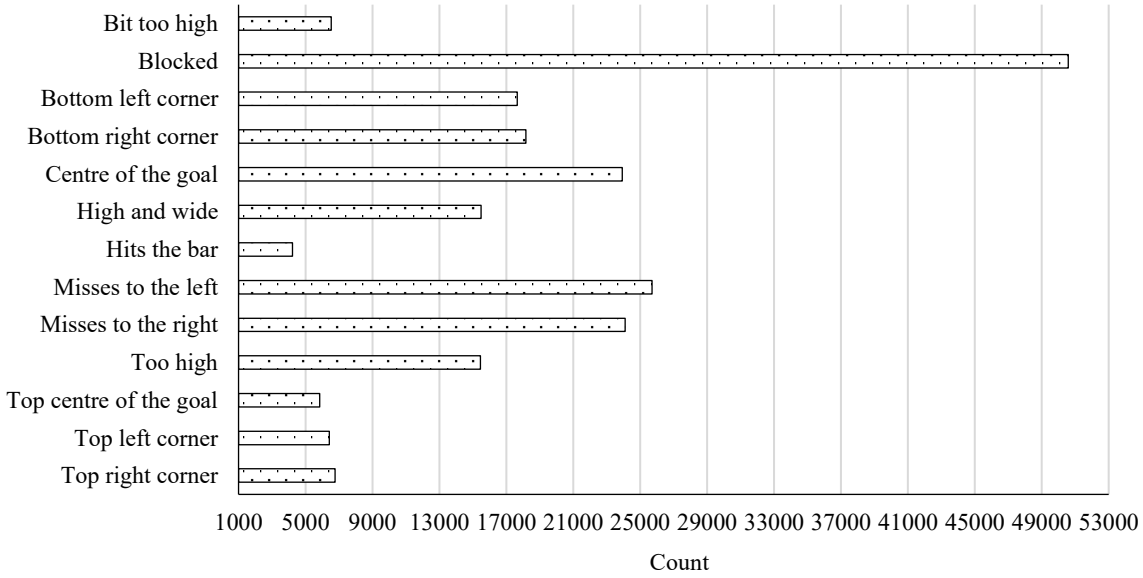

Figure 3. Shot placement.

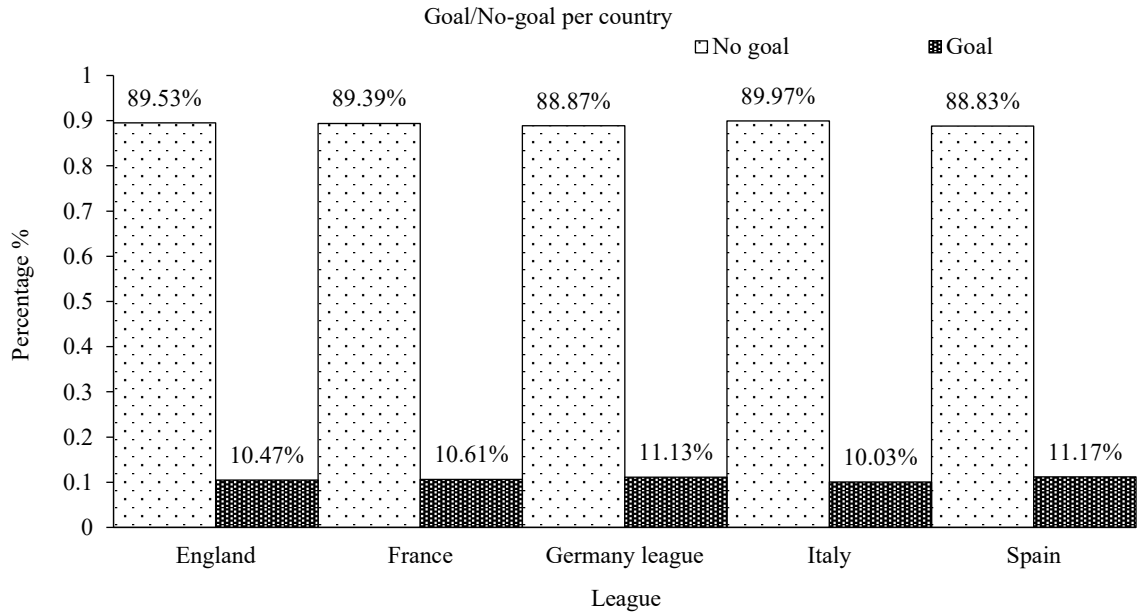

Figure 4. Goal/No-goal per country.

As illustrated in Figure 4, minimal changes are observed across various major leagues. It seems that any given attempt has a 10–11% chance of success universally.

As seen in Figure 5, the goal/no goal percentages remain practically constant throughout time. So, it is becoming evident that, statistically speaking, one out of every nine to ten shots is a goal, regardless of where or when you look.

As shown in Figure 6 and Table 1, our xG model can properly identify whether a shot is a goal 91% of the time. Furthermore, we get a ROC-AUC measure of 82%. This appears to be quite promising. However, these two criteria fail to account for our dataset's extreme imbalance [14]. There are significantly more unsuccessful shots than successful ones. Therefore, for instance, predicting that every shot will fail to score would yield an accuracy rate of 89%. Hence, additional metrics are required to assess the effectiveness of our model.

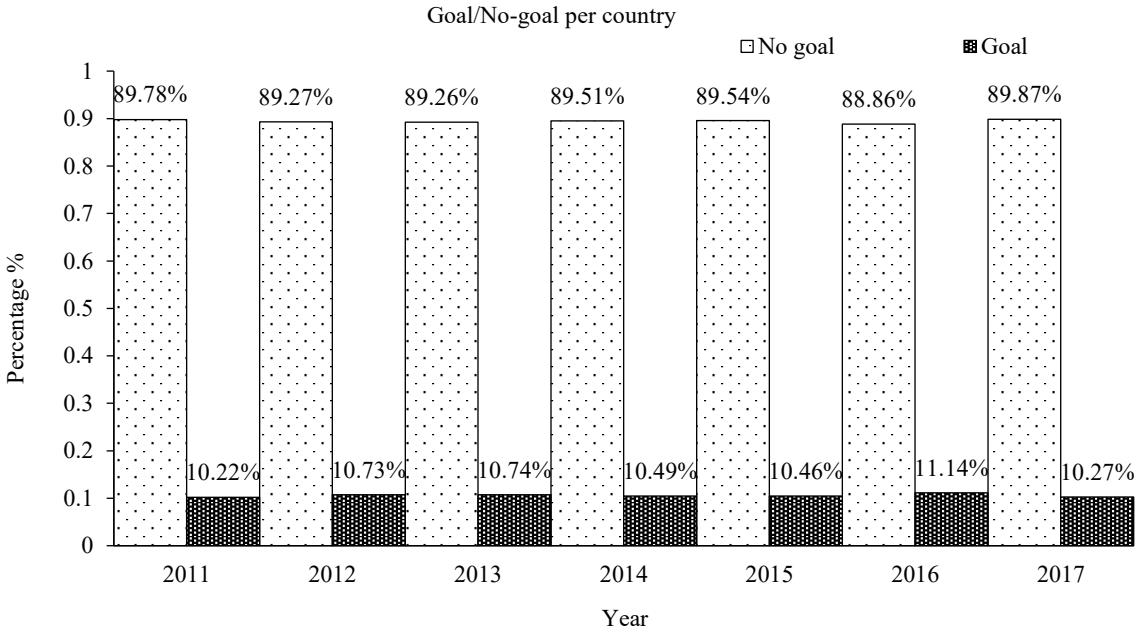

Figure 5. Goal/No-goal per year.

Table 1. Model performance metrics for different parameter configurations.

f1_score	0.390152	0.390015	0.389826	0.389793	0.389761
learning_rate	0.285508	0.247483	0.118222	0.131184	0.262329
loss	–0.818621	–0.818853	–0.819267	–0.819361	–0.818056
max_depth	19	6	5	5	15
max_features	7	21	24	19	5
min_samples_leaf	99	80	149	143	17
precision	0.714465	0.714375	0.717302	0.717076	0.710199
recall	0.268344	0.268227	0.267639	0.267639	0.268579
status	ok	ok	ok	ok	ok
test_ROCAUC	0.818621	0.818853	0.819267	0.819361	0.818056
test_accuracy	0.911045	0.911033	0.911157	0.911145	0.910821
train_R	0.8188	0.8182	0.8175	0.8175	0.8195

```
GradientBoostingClassifier (criterion='friedman_mse', init=None,
                learning_rate=0.285508, loss='deviance', max_depth=19,
                max_features=7, max_leaf_nodes=None,
                min_impurity_decrease=0.0, min_impurity_split=None,
                min_samples_leaf=99, min_samples_split=2,
                min_weight_fraction_leaf=0.0, n_estimators=100,
                n_iter_no_change=None, presort='auto', random_state=None,
                subsample=1.0, tol=0.0001, validation_fraction=0.1,
                verbose=0, warm_start=False)

 The test set contains 80198 examples (shots) of which 8504 are positive (goals).
 The accuracy of classifying whether a shot is goal or not is 91.0%.
 Our classifier obtains an ROC-AUC of 82.0%.
 The baseline performance for PR-AUC is 8.11%. This is the PR-AUC that what we would
 get by random guessing.
 Our model obtains an PR-AUC of 47.37%.
 Our classifier obtains a Cohen Kappa of 8.35.
```

Figure 6. Summary of model parameters and evaluation metrics.

As demonstrates in the Figure 7 how the confusion matrix summarizes all predictions. Our model correctly identified 70,781 shots that did not hit the target, but was wrong in 6,238 cases where it predicted the shot wouldn't hit the target, but it did. Another column shows that it correctly predicted 913 targets but failed to predict 2266 shots as targets. The analysis indicates that the model excels at predicting class 0 (no-goal), but struggles to forecast class 1 (goals) accurately. The latter achieves 71% precision and 27% recall, yielding an F1 score of 0.39. These numbers are fair, but not exceptional. In Table 2, we see the rank, player, true goals and expected goals.

Messi outperforms all other players on this criteria as shown in the Figure 8. He scored 205 goals, exceeding the projected 146 based on the number and qualities of his shots. The top of the ranking is dominated by world-class players [14]. This section focuses on absolute values, or overall aims. Consider the ratio of goals scored to projected goals. It will only include players who scored more than 30 goals over an 8-year period (a modest number).

Confusion matrix:
[[70781 913]
[6238 2266]]
Report:

	precision	recall	f1-score	support
0	0.92	0.99	0.95	71694
1	0.71	0.27	0.39	8504
micro avg	0.91	0.91	0.91	80198
micro avg	0.82	0.63	0.67	80198
weighted avg	0. 98	0. 91	0. 89	80198

Figure 7. Confusion matrix.

Table 2. Top players ranked by difference between true and expected goals.

Rank	Player	Difference	True goals	Expected goals
1	Lionel Messi	–58.85	205	146.15
2	Zlatan Ibrahimovic	–33.77	153	119.23
3	Cristiano Ronaldo	–32.25	198	165.75
4	Gonzalo Higuain	–31.69	118	86.31
5	Luis Suarez	–31.68	96	64.32

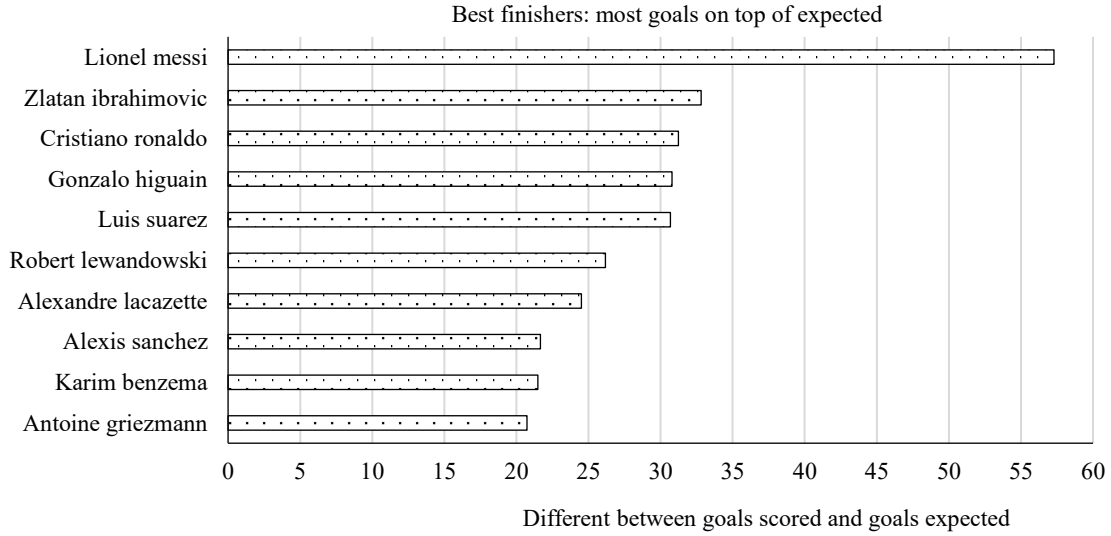

Figure 8. Different between goals scored and goals expected.

As Figure 9 shows Ribery is the game's most prolific finisher. This measure is simple and straightforward to interpret [14]. The proportion of goals scored by a player relative to the average player's shot execution is referred to as their scoring ratio. Ribery scored nearly twice as many goals as expected based on his shot count and context. Table 3 shows the league data and goal/xgoals Ratio.

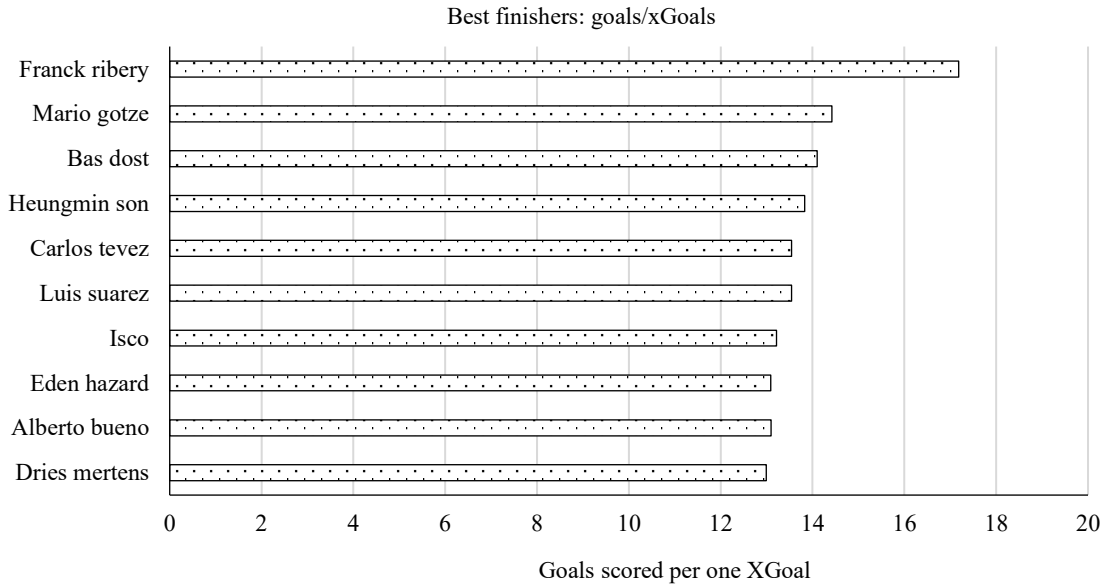

Figure 9. Goals scored per one XGoal.

Table 3. League table goals/xgoals ratio.

League	Year	Best finisher	Goals	Goals/Goals Ratio
France	2011	Olivier Giroud	13	1.23
France	2012	Zlatan Ibrahimovic	18	1.47
France	2013	Dario Cvitanich	16	1.69
France	2014	Cheick Diabaté	13	1.61
France	2015	Benjamin Moukandjo	13	1.89
France	2016	Zlatan Ibrahimovic	21	2.26
Germany	2011	Klaasjan Huntelaar	15	1.81
Germany	2012	Martin Harnik	15	1.79
Germany	2013	Ivica Olic	13	2.22
Germany	2014	Josip Drmic	14	2.03
Germany	2015	Raffael	13	1.81
Germany	2016	Pierre-Emerick Aubameyang	23	1.34
Italy	2012	Miroslav Klose	14	1.99
Italy	2013	Luis Muriel	13	1.62
Italy	2014	Ciro Immobile	13	1.78
Italy	2015	Massimo Maccarone	15	1.79
Italy	2016	Dries Mertens	15	1.79
Spain	2011	Lionel Messi	17	1.55
Spain	2012	Lionel Messi	59	1.58
Spain	2013	Pedro	16	2.01
Spain	2014	Cristiano Ronaldo	34	1.59
Spain	2015	Antoine Griezmann	25	1.77
Spain	2016	Gareth Bale	16	1.81
England	2013	Luis Suarez	16	1.83
England	2014	Christian Eriksen	13	1.88
England	2015	Odion Ighalo	14	1.74
England	2016	Eden Hazard	13	1.78

As Shown in the Figure 10, it encounter players who score fewer goals than expected, such as Mario Balotelli, Giampaolo Pazzini or Edin Dzeko. Giampaolo Pazzini is featured on both lists, highlighting the disparity between actual and expected goals/xGoals ratio [14]. Figure 11 shows the highest value of total goals vs. Total xGoals across all seasons by player.

As displayed in Table 4, Tom Huddlestone seems to be having trouble deciding when to shoot as the expected shot on target is 0.03. Notably, instances involving Gohhan Inler and Ruben Rochina are intriguing. Despite frequently attempting unconventional shots from long distances, they seem to excel at them. Figure 12 shows that player should have best shot decider vs xG value per shot.

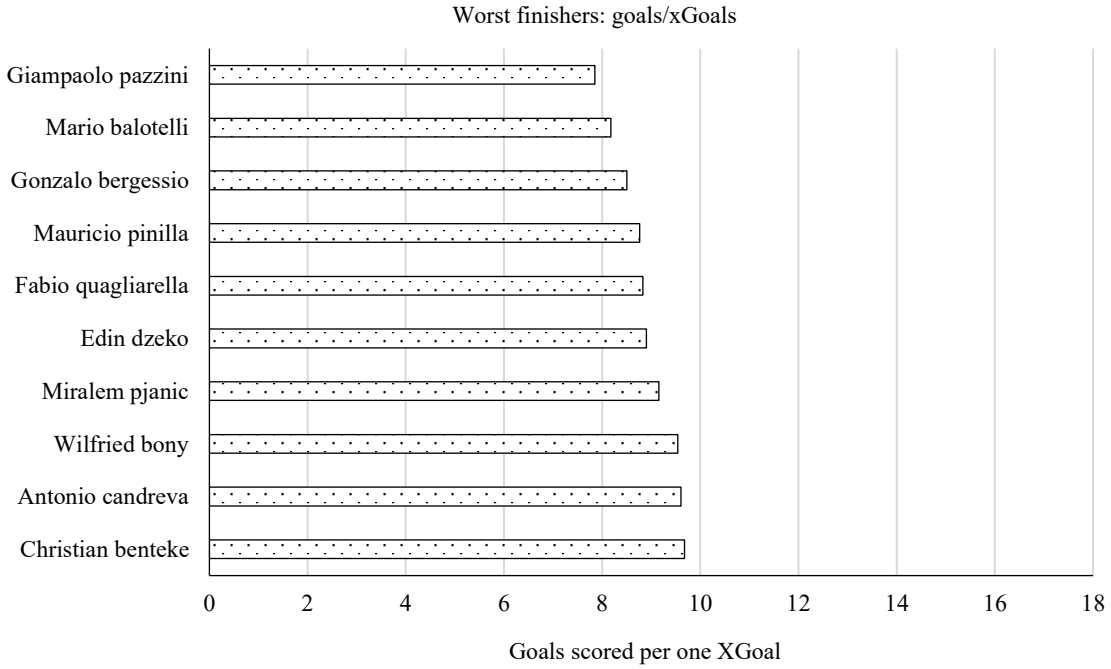

Figure 10. Worst finishers: goals/xgoals.

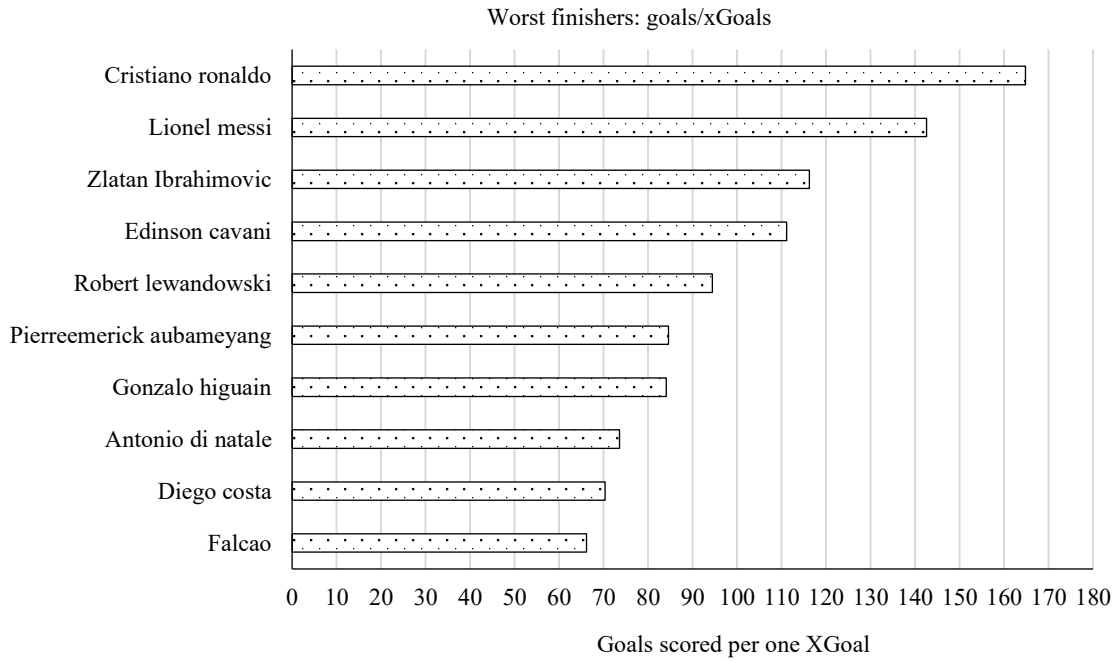

Figure 11. Highest values of total xgoals.

Table 4. Player ranking by xG per shot, goals, expected goals, and difference.

Rank	Player	xG_per_shot_ratio	True Goals	Expected Goals	difference
670	Daniel Baier	0.042260	5	7.48	2.48
671	Ivan Radovanovic	0.040994	2	6.60	4.60
672	Florent Balmont	0.036786	5	6.18	1.18
673	Gokhan Inler	0.035730	9	6.61	-2.39
674	Tom Huddlestone	0.030463	2	3.29	1.29

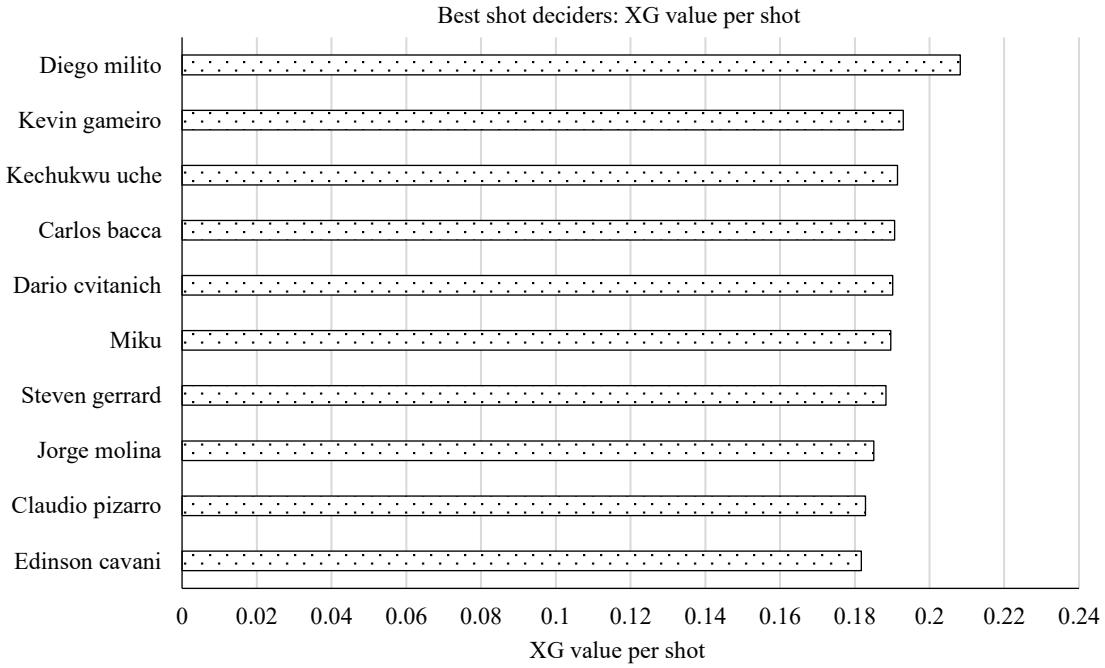

Figure 12. Best shot deciders: xG value per shot.

Table 5. Surpassing expectations with his left-footed goals.

Rank	Player	n_leftFoot_shots	True Goals	Expected Goals	Difference
1	Lionel Messi	752	167	121.593427	–45.406573
2	Antoine Griezmann	345	58	41.405059	–16.594941
3	Arjen Robben	296	42	32.100030	–9.899970
4	Iago Falque	132	23	13.249413	–9.750587
5	Franck Ribery	57	16	6.454551	–9.545449

Table 6. Exceeds expectations with his left-footed goals.

Rank	Player	Expected Goals	True Goals
80	Cristiano Ronaldo	28.452525	32

As shown in Table 5, Messi leads the list, surpassing expectations with his left-footed goals. Antoine Griezmann, Iago Falque, and Arjen Robben follow suit, showcasing their prowess with their left foot.

As illustrated in Table 6, he still exceeds expectations with his left-footed goals, indicating his exceptional ability with that foot. However, he does not rank among the top players in this regard.

Table 7 shows that new players such as Mohamed Salah, who became a professional left-foot finisher. As shown in the Table 8, it depicts Luis Suarez holds the record for his ability to convert shots into goals.

As shown in the Table 9, it offers an interesting comparison of goals from different players, including Messi, Cristiano Ronaldo, Zlatan Ibrahimovic and Robert Lewandowski.

As illustrated in Figure 13, it shows comparison between headers, right foot and left foot. Ronaldo emerges as the best header, while Lewandowski trails behind. Messi excels with his left foot, while Zlatan appears dominant with his right.

Table 7. Professional left-foot finisher.

Rank	Player	n_leftFoot_shots	True Goals	Expected Goals	Ratio
1	Iago Falque	132	23	13.249413	1.735926
2	Mohamed Salah	129	23	13.899066	1.654787
3	Lukas Podolski	132	21	13.196789	1.591296
4	James Rodriguez	162	23	15.533109	1.480708
5	Zlatan Ibrahimovic	129	25	17.293942	1.445593

Table 8. Top right-footed shooters: shots, goals, expected goals, and difference.

Rank	Player	n_rightFoot_shots	True Goals	Expected Goals	Difference
1	Luis Suarez	289	69	43.197401	–25.802599
2	Gonzalo Higuain	362	86	60.853048	–25.146952
3	Alexandre Lacazette	270	70	48.149617	–21.850383
4	Zlatan Ibrahimovic	554	111	90.985123	–20.014877
5	Robert Lewandowski	378	84	65.414378	–18.585622

Table 9. Player ranking by right foot shots, goals, expected goals, and scoring ratio.

Rank	Player	n_rightFoot_shots	True Goals	Expected Goals	Ratio
1	Bas Dost	62	23	13.025501	1.765767
2	Carlos Tevez	181	33	19.852676	1.662244
3	Franck Ribery	165	24	14.466810	1.658970
4	Mario Gotze	170	32	19.853448	1.611811
5	Lionel Messi	109	30	18.746928	1.600262

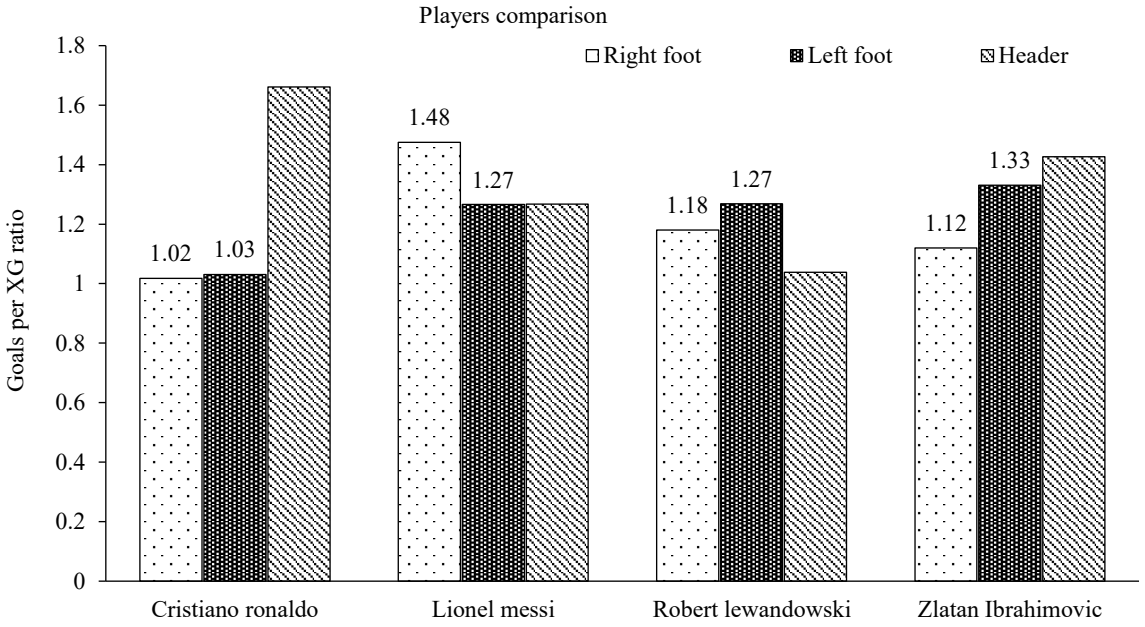

Figure 13. Comparison between headers, right foot and left foot.

Figure 14 illustrates that while random players can achieve high values (e.g., Franck Dja Djedje's right foot ratio is higher than the top four), their consistency is low, with many values falling below 1.0. It is likely that the high values are due to a lack of shots, as a single goal can significantly impact the measure.

As seen in Table 10, Messi, Pogba, and Zlatan outshine their predicted output in terms of goals scored from outside the box. Conversely, players like Mario Balotelli and Alessandro Diamanti underperform in this aspect. As shown in the Table 11, James Rodriguez and Yaya Toure have maximum outbox shots. As shown in the Figure 15, James Rodriguez is the best shooter from outside the box in the game.

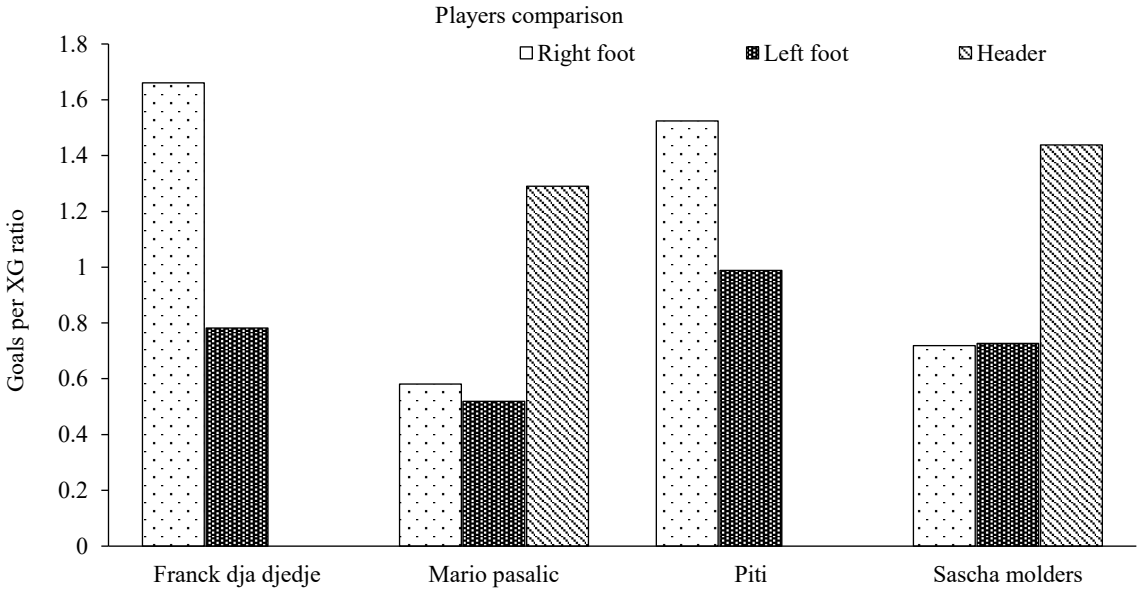

Figure 14. Random players can achieve high values.

Table 10. Goals scored from outside the box.

Rank	Player	n_outbox_shots	True Goals	Expected Goals	Difference
1	Lionel Messi	304	16	6.373508	–9.626492
2	Paul Pogba	226	14	5.983467	–8.016533
3	Zlatan Ibrahimovic	261	14	6.775687	–7.224313
4	Gonzalo Higuain	127	11	3.856737	–7.143263
5	Yaya Toure	128	10	3.380151	–6.619849
4144	Alberto Aquilani	117	0	3.146898	3.146898
4145	Ronny Rodelin	119	0	3.204116	3.204116
4146	Francesco Lodi	170	0	3.445538	3.445538
4147	Alessandro Diamanti	254	3	6.506579	3.506579
4148	Mario Balotelli	209	1	4.735436	3.735436

Table 11. Maximum outbox shots.

Rank	Player	n_outbox_shots	True Goals	Expected Goals	Ratio
1	James Rodriguez	105	9	2.422328	3.715434
2	Alain Traore	86	6	1.832104	3.274923
3	Julian Draxler	94	8	2.609330	3.065921
4	Alexandre Lacazette	100	9	2.951516	3.049280
5	Yaya Toure	128	10	3.380151	2.958447

As shown in Table 12, Messi is at the top. Over a 7-year period, this player's passing has resulted in the most predicted goals for his teammates.

As shown in the Table 13, Messi had the maximum number of the passes. After him, Angel is in the second position on the number of the passes.

As shown in the Figure 16, this squad includes players such as Luis Suarez, Angel Di Maria and Gareth Bale, who put in dangerous performances with goals in high demand.

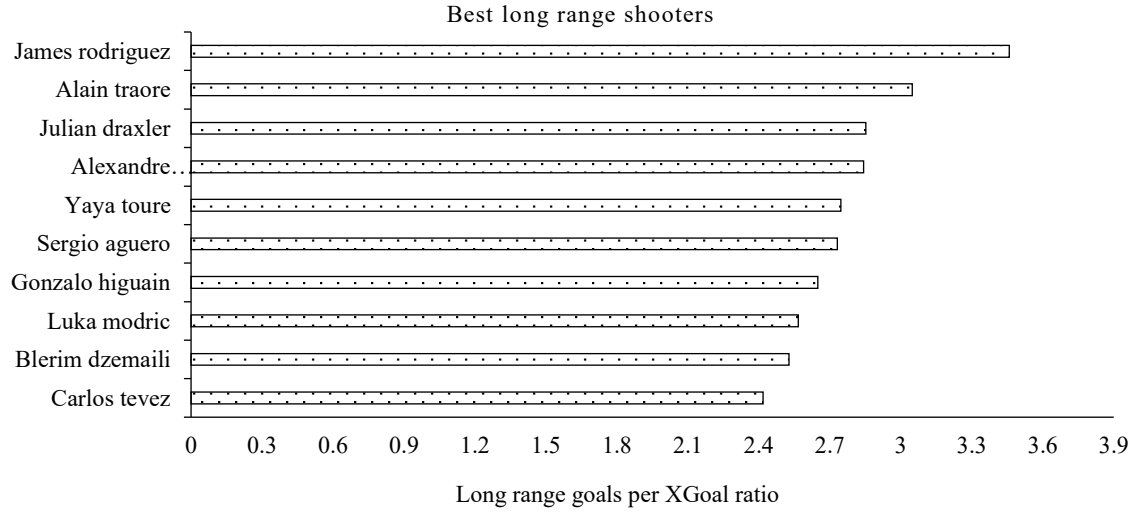

Figure 15. Long range shooters.

Table 12. Top players ranked by passes and expected goals created from passing.

Rank	Player	n_passes	trueGoals_created	expectedGoals_created
1	Lionel Messi	350	68	51.133392
2	Mesut Ozil	343	35	36.899926
3	Cesc Fabregas	264	53	36.468663
4	Zlatan Ibrahimovic	270	36	36.231185
5	Marek Hamsik	370	50	34.910338

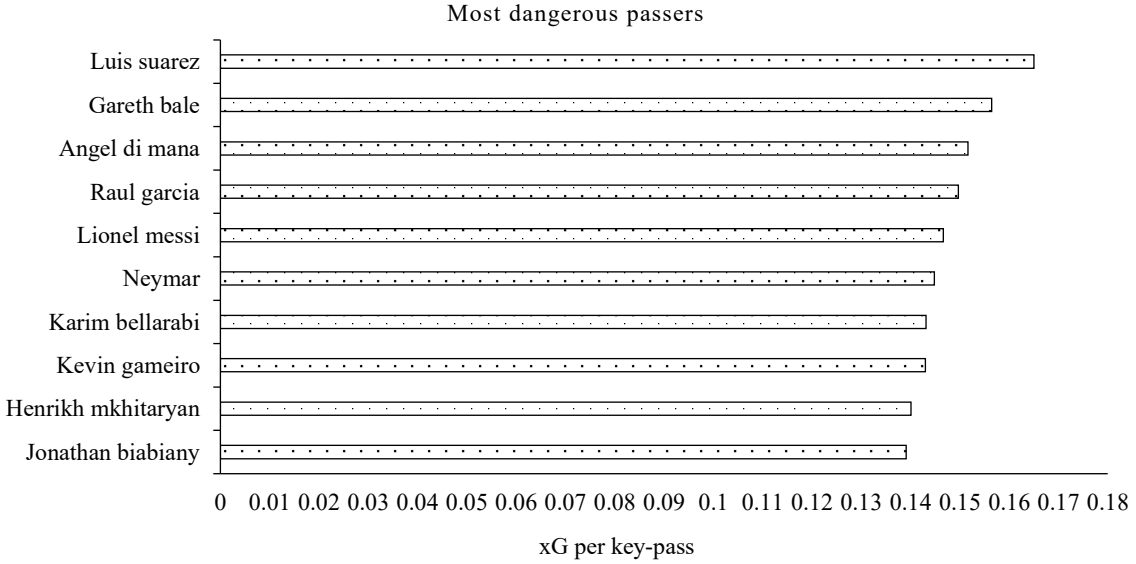

Figure 16. Most dangerous passers.

Table 13. Top players ranked by number of passes and xG per pass.

Rank	Player	n_passes	xG_perpass
1	Luis Suarez	185	0.164643
2	Gareth Bale	109	0.156140
3	Angel Di Maria	211	0.151309
4	Raul Garcia	90	0.149269
5	Lionel Messi	350	0.146095

Table 14. Excepted goals passes and true goals.

Rank	Player	n_passes	trueGoals_created	expectedGoals_created	difference
1	Joan Verdu	134	7	14.545416	7.545416
2	Xabi Prieto	152	6	13.199252	7.199252
3	Philippe Coutinho	177	13	19.910849	6.910849
4	Luca Cigarini	164	6	12.901295	6.901295
5	Alejandro Gomez	137	5	11.747375	6.747375
687	Marek Hamsik	370	50	34.910338	-15.089662
688	Cristiano Ronaldo	222	45	29.263254	-15.736746
689	Cesc Fabregas	264	53	36.468663	-16.531337
690	Karim Benzema	214	40	23.423395	-16.576605
691	Lionel Messi	350	68	51.133392	-16.866608

As seen in Table 14, it compares players on the basis of the excepted goals passes and true goals. The players from top-tier teams dominate the list, while notable players like Philippe Coutinho and Eden Hazard seem to be let down by their teammates' finishing abilities.

CONCLUSION

In the conclusion, the expected Goals (xG) is a statistical metric used in football (soccer) to assess the quality of goal-scoring opportunities created by a team or player during a match. It provides a numerical value to quantify the likelihood of a shot resulting in a goal based on various factors such as shot distance, angle, assist type, and other situational variables. Overall, xG provides valuable insights into player and team performance beyond just the number of goals scored, allowing coaches, analysts, and fans to understand the underlying quality of scoring opportunities and make informed decisions about tactics, player selection, and performance evaluation. In this we can see that our xG model is able to correctly predict whether a shot is goal or not 91% of the times. Furthermore, we obtain a pretty good ROC-AUC metric of 82%. We can compile a confusion matrix to encapsulate all predictions, revealing that our model accurately identifies 70,781 shots as non-goals but misjudges 6,238 instances where it predicts a shot will not result in a goal, yet it does. Conversely, the other column correctly predicts 913 goals but fails to anticipate 2266 successful shots as goals. The report highlights the model's proficiency in predicting class 0 (no-goal) but its struggle to forecast class 1 (goals), with the latter exhibiting 71% accuracy and 27% recall, yielding an F1 score of 0.39. While these metrics are commendable, they fall short of exemplary standards.

REFERENCES

1. Wang Z. Semantic analysis based on fusion of audio/visual features for soccer video. Procedia Comput Sci. 2021 Jan 1; 183: 563–71.
2. Zhou J, Shen X, Wang J, Zhang J, Sun W, Zhang J, Birchfield S, Guo D, Kong L, Wang M, Zhong Y. Audio-visual segmentation with semantics. Int J Comput Vis. 2025 Apr; 133(4): 1644–64.
3. Oskouie P, Alipour S, Eftekhari-Moghadam AM. Multimodal feature extraction and fusion for semantic mining of soccer video: a survey. Artif Intell Rev. 2014 Aug; 42(2): 173–210.

4. Qian X, Wang H, Liu G, Hou X. HMM based soccer video event detection using enhanced mid-level semantic. Multimed Tools Appl. 2012 Sep; 60(1): 233–55.
5. Tjondronegoro DW, Chen YP. Knowledge-discounted event detection in sports video. IEEE Trans Syst Man Cybern-Part A: Syst Hum. 2010 May 20; 40(5): 1009–24.
6. Xu C, Wang J, Lu H, Zhang Y. A novel framework for semantic annotation and personalized retrieval of sports video. IEEE Trans Multimed. 2008 Mar 21; 10(3): 421–36.
7. Yu J, Lei A, Hu Y. Soccer video event detection based on deep learning. In International Conference on Multimedia Modeling. Cham: Springer International Publishing; 2018 Dec 11; 377–389.
8. Xu M, Xu C, Duan L, Jin JS, Luo S. Audio keywords generation for sports video analysis. ACM Trans Multimed Comput Commun Appl. 2008 May 16; 4(2): 1–23.
9. Zhang YZ, Wang JY, Dai YW. Soccer video shot segmentation based on self-adapting dual threshold and dominant color percentage. J Nanjing Univ Sci Technol (Nat Sci). 2009; 33(4): 432–7.
10. Yu JQ, Wang N. Shot classification for soccer video based on sub-window region. J Image Graph. 2008; 13(7): 1347–1352.
11. Huang Q, Hu J, Hu W, Wang T, Bai H, Zhang Y. A reliable logo and replay detector for sports video. In 2007 IEEE International Conference on Multimedia and Expo. 2007 Jul 2; 1695–1698.
12. Yu JQ, He HH, He YF. Highlights extraction for soccer video based on affection arousal. J Comput Res Dev. 2010; 47(10): 1823–31.
13. Hanjalic A. Adaptive extraction of highlights from a sport video based on excitement modeling. IEEE Trans Multimed. 2005 Nov 21; 7(6): 1114–22.
14. Gabriel Manfredi. (2020). Expected Goals & Player Analysis. [Online]. Kaggle. Available from: https://www.kaggle.com/code/gabrielmanfredi/expected-goals-player-analysis

