# A Comprehensive Framework for Predictive Modeling of Football Match Outcomes in Germany's 3. Liga

## Part I: Foundational Framework for Football Prediction

### Section 1: The Analytical Challenge of Football: A Game of Low Numbers and High Variance

The pursuit of predicting football match outcomes through statistical and machine learning models is a sophisticated endeavor, fraught with challenges unique to the sport. Unlike high-scoring games such as basketball, football is fundamentally a game of low numbers, a characteristic that profoundly influences its predictability.[1] A typical match may see only two or three goals in total, meaning that the final score is highly susceptible to the influence of random events. A single deflected shot, a controversial refereeing decision, or a moment of individual brilliance or error can be the sole determinant of a match's outcome, swinging the result from a draw to a win or a loss. This high degree of variance means that the final result is often a poor representation of the underlying performance balance between the two competing teams.

This inherent randomness finds a natural mathematical expression in the Poisson distribution. A large body of academic research has established that the number of goals scored by a team in a football match can be reasonably approximated by a Poisson process, where events (goals) occur randomly and independently at a known average rate.[3] This statistical reality serves as the cornerstone for many foundational predictive models and underscores a critical principle: any successful model must account for, and attempt to see through, the noise of stochasticity. The final scoreline, while definitive, is an extremely noisy signal of which team was truly superior during the 90 minutes of play. A team can dominate every phase of the game, creating numerous high-quality scoring opportunities, yet lose 1-0 due to a single counter-attack. Conversely, a team can be thoroughly outplayed but win through a fortunate goal.

Consequently, models trained exclusively on the final match outcome (Win, Draw, Loss) will inevitably struggle to learn the true, underlying strength of the teams. The model is being fed a noisy target variable, making it difficult to separate the signal of genuine team quality from the noise of in-game luck. The core analytical challenge, therefore, is not necessarily a lack of data, but the difficulty of extracting a clear and stable signal of performance from a data stream dominated by noise. The solution lies not merely in accumulating more data, but in engineering better features and employing metrics that measure the *process* of performance rather than just the final, often random, outcome.[2]

Furthermore, the three-way outcome of a football match (Home Win, Draw, Away Win) presents a distinct modeling challenge. It is a multi-class classification problem where one of the classes, the draw, is often the most difficult to predict accurately. Draws can be under-represented in datasets and arise from complex game dynamics that are not easily captured by simple metrics, posing a significant hurdle for many predictive algorithms.[7]

Addressing these fundamental characteristics—the low-scoring nature, the disconnect between performance and results, and the difficulty of predicting draws—is the essential first step in constructing a robust and meaningful predictive system.

## Section 2: Data Infrastructure for the 3. Liga Project

Building a successful predictive model begins with the construction of a comprehensive and reliable data infrastructure. For a project focused on a lower-tier league like Germany's 3. Liga, this phase is particularly critical, as data is often less centralized and less detailed than for top-flight competitions. A successful approach requires a multi-source strategy to collate a rich dataset.

### Sourcing and Acquiring Data for the 3. Liga

No single data provider is likely to offer all the necessary data for a sophisticated model of a third-tier league for free. Therefore, a hybrid data strategy is mandated, fusing information from various sources to build a complete picture of each match.

- **Free Web Sources for Foundational Data:** A wealth of historical data, including match results, league tables, and basic statistics, is available from publicly accessible websites. These form the bedrock of the dataset. Key sources include **FBref** [10],
**TheSportsDB** [11],
**Soccer24** [12],
**Soccerway** [13], and
**Flashscore**.[14] Of these, FBref is particularly noted for its statistical depth, while TheSportsDB offers a free API that includes historical league winners and event data stretching back several seasons, making it a valuable resource for programmatic

access.[10]

- **APIs for Programmatic Access:** For more structured and automated data collection, several APIs provide coverage for the 3. Liga. The **football-data.org** API explicitly lists Germany's 3. Liga under the code "BL3" and provides endpoints for matches, teams, standings, and top scorers for both current and past seasons.[15] This is an excellent starting point for building an automated data pipeline.
- **Paid/Freemium APIs for Advanced Data:** For more advanced features, paid services are often necessary. Providers like **Sportmonks** [16] and **Live-score-api** [17] offer more granular data, including detailed in-match statistics, live scores, and, crucially, betting odds. While this project may aim to start with free data, awareness of these services is important for future expansion.
- **Community-Sourced Datasets:** Platforms like **Kaggle** host numerous football datasets. While many focus on top leagues [18], some are more comprehensive. The "German Football Scores" dataset, for instance, is noted to contain liga_3.json files, suggesting it holds historical fixture and ranking data relevant to this project.[19] Exploring such datasets can provide a valuable, pre-cleaned starting point.
- **Specialized Data:** For certain features, specialized sources are invaluable. **Transfermarkt**, for example, provides detailed information on team and player market values.[20] While not a traditional performance metric, squad market value can serve as a powerful proxy for team quality and financial strength.

The data landscape for the 3. Liga reveals a crucial reality: while basic results are plentiful, advanced performance metrics like Expected Goals (xG) are not as readily available as they are for the Premier League or Bundesliga. This "lower-league data deficit" elevates the importance of other data types. In this context, betting odds are not merely an optional feature for a financial simulation; they become a critical *proxy* for the sophisticated information that is otherwise missing. Bookmakers' odds encapsulate a vast array of inputs, including team news, expert analysis, and proprietary performance models.[21] By incorporating odds, the model can gain access to a distilled form of this high-level intelligence, partially compensating for the absence of granular event data. This necessitates a data engineering pipeline capable of scraping basic results from one source, accessing odds via an API from another, and potentially pulling squad values from a third, before meticulously cleaning and merging these disparate datasets into a unified whole.

## Data Acquisition and Structuring

Given that many free sources lack a formal API, web scraping becomes an indispensable skill. The process generally involves using Python libraries such as requests to fetch the HTML content of a webpage and BeautifulSoup or lxml to parse it, extracting data from the underlying table structures.[22]

Once acquired, the raw data requires significant cleaning and structuring. This pre-processing phase includes:

1. **Standardizing Team Names:** Different sources may use slightly different names for the same team (e.g., "1860 Munich" vs. "TSV 1860 München"). A mapping file or function is needed to ensure consistency.
2. **Converting Data Types:** Dates, scores, and other numerical data must be converted from strings to the appropriate datetime or numeric types for analysis.
3. **Creating a Match-Centric Structure:** The final dataset should be structured with one row per match, containing columns for the date, home team, away team, full-time score, and all the engineered features that will be developed in the next section.

The following table provides a summary of potential data sources to guide this acquisition process.

**Table 1: Comparison of Data Sources for Germany's 3. Liga**

| Source Name | URL | Data Types Covered | Access Method | Historical Depth | Cost | Key Considerations |
|---|---|---|---|---|---|---|
| **FBref** | fbref.com | Match Results, Basic & Advanced Stats (xG may be available) | Web Scrape | Extensive, since 2018 launch [10] | Free | Excellent source for detailed team and player stats. Data is well-structured in HTML tables. |
| **TheSportsDB** | thesportsdb.com | Match Results, Fixtures, Team Info, Historical Winners | Free API | First recorded event 2019 [11] | Free | Provides a user-friendly API, good for automating basic data collection. |
| **football-data.org** | football-data.org | Matches, Teams, Standings, Scorers, Odds | Free/Paid API | Multiple past seasons available [15] | Freemium | Excellent API with a dedicated endpoint for 3. Liga ("BL3"). Free tier has limitations. |
| **Soccer24** | soccer24.com | Match Results, Fixtures, Standings, | Web Scrape | Extensive archive [12] | Free | Good for cross-referencing results and |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | Archive | | | | historical league tables. |
| **Sportmonks** | sportmonks.com | Live Scores, Detailed Stats, Fixtures, Odds | Paid API | Over 1,300 leagues covered [16] | Paid | Professional-grade data, likely includes rich features like xG but comes at a cost. |
| **Transfermarkt** | transfermarkt.us | Player/Team Market Values, Transfer History | Web Scrape | Extensive [20] | Free | The definitive source for market value data, a powerful proxy for team quality. |
| **Kaggle** | kaggle.com | Various (Fixtures, Rankings, Results) | CSV/JSON/SQL | Varies by dataset | Free | Can provide large, pre-cleaned datasets, but 3. Liga specific data needs to be verified.[19] |

## Section 3: The Science of Feature Engineering: From Raw Data to Predictive Power

The single most impactful phase of a predictive modeling project is feature engineering. The quality and predictive power of the features will have a greater influence on the final model's performance than the specific choice of algorithm. A sophisticated model fed with naive features will underperform a simpler model fed with a rich, well-engineered feature set. The goal is to transform the raw, collated data from Section 2 into a set of variables that capture the multifaceted nature of football performance. A robust feature set should not be redundant; instead, it should combine features that measure different, complementary aspects of a team's capabilities. These can be broadly categorized into four pillars: recent outcomes (Form), underlying process quality (Advanced Metrics), long-term historical strength (Rating Systems), and collective market expectation (Market Intelligence).

### 3.1. Foundational Statistical Features: Capturing Form and Context

These features are derived from basic match statistics and form the initial layer of predictive information.

- **Form and Momentum:** The most intuitive features relate to a team's recent performance. This is typically calculated using rolling averages over a defined window of past matches (e.g., the last 5, 10, or 20 games). Key features include:
  - Points Per Game (PPG) over the window.
  - Average Goals Scored (GS) per game.
  - Average Goals Conceded (GC) per game.
  - Average Goal Difference (GD) per game.
    The concept of using past match data to inform future predictions is a fundamental starting point for any sports model.22
- **Home vs. Away Dichotomy:** Home advantage is a persistent and significant phenomenon in football, influenced by factors like crowd support, familiarity with the pitch, and travel fatigue for the opposition.[4] It is therefore critical to disaggregate performance data. Instead of a single "Goals Scored" feature, a model should be fed separate
  HomeGoalsScored_form and AwayGoalsScored_form features, calculated exclusively from a team's past home or away matches, respectively. This provides the model with crucial context about how a team performs under different conditions.

### 3.2. Advanced Performance Metrics: The Power of Expected Goals (xG)

To move beyond the noisy signal of actual goals, it is necessary to use metrics that evaluate the quality of the performance process itself. The most powerful and widely adopted metric for this is Expected Goals (xG).

- **The Concept of xG:** Expected Goals measures the quality of a scoring opportunity. It assigns every shot a probabilistic value, from 0 to 1, representing the likelihood of that shot resulting in a goal based on a historical analysis of thousands of similar shots.[25] A penalty kick, for instance, has an xG of approximately 0.76, as penalties are historically converted 76% of the time.[29] The power of xG lies in its ability to separate finishing skill (or luck) from the process of chance creation. Over the long term, a team's actual goals tend to converge towards their xG, making xG a more stable and predictive indicator of future performance than raw goal counts.[2]
- **Sourcing or Modeling xG:** For the 3. Liga, pre-calculated xG data may be difficult to source for free. Two paths are available:
  1. **Sourcing:** Investigate paid data providers like Sportmonks or specialized analytics companies that may offer xG data for lower-tier German leagues.

2. **Modeling:** A simplified xG model can be built in-house. This typically involves a logistic regression or a machine learning model like XGBoost trained on shot event data.[31] The necessary features for such a model would include, at a minimum, the shot's location (distance and angle to goal). More advanced models incorporate features like the body part used (head vs. foot), the pattern of play (open play, set piece, fast break), and defensive pressure.[25] While event-level data for the 3. Liga may be sparse, even a simple location-based xG model is superior to using raw shot counts.

- **xG as a Predictive Feature:** Once xG values for each shot are obtained, they can be aggregated at the match level and used to create exceptionally powerful predictive features. These are typically calculated over a rolling window, just like basic form metrics [32]:
    - xG_for (xGF): Average Expected Goals created per game.
    - xG_against (xGA): Average Expected Goals conceded per game.
    - xG_difference (xGD): The difference between xGF and xGA.
    - xG_ratio (xGR): Calculated as xGF / (xGF + xGA), this metric, highlighted in some analyses, provides a normalized measure of dominance.[32]

### 3.3. Dynamic Team Strength: The Elo Rating System

While form metrics capture recent performance, they can be volatile. A single measure of a team's long-term, underlying strength is needed to provide historical context. The Elo rating system is perfectly suited for this.

- **The Concept of Elo:** Originally developed for chess, Elo is a zero-sum rating system. After every match, the winning team takes points from the losing team. The number of points exchanged depends on the probability of the outcome, which is derived from the difference in the teams' ratings before the match.[7] A surprise victory by a low-rated team over a high-rated team results in a large exchange of points, while an expected victory by a strong team over a weak one results in a small exchange. This creates a single, dynamic rating for each team that reflects its current strength relative to all other teams in the league.
- **Implementation as a Feature:** The Elo system can be implemented with a simple script. The key inputs are the initial ratings for each team at the start of the season, the match results, and a "K-factor" that controls how much the ratings change after each match. An adjustment can also be added to account for home-field advantage. The resulting Elo ratings for the home and away teams, and the difference between them, serve as powerful features that capture the long-term class and trajectory of the teams involved.

### 3.4. Market Intelligence: Incorporating Betting Odds

Betting odds are a unique and invaluable source of information. They represent the aggregated prediction of a highly sophisticated and financially incentivized market.[34] The odds offered by bookmakers reflect not only quantitative statistical models but also a vast amount of qualitative information that is difficult to model otherwise, such as last-minute injuries, team morale, or tactical changes.[21]

- **Odds as Predictive Features:** The closing odds for the three outcomes (Home Win, Draw, Away Win) can be converted into "implied probabilities." This requires removing the bookmaker's margin (the "overround" or "vig"), which ensures the sum of the probabilities is greater than 100%, guaranteeing the bookmaker's profit. Once the margin is removed, these implied probabilities serve as extremely strong baseline predictors.[8]
- **Advanced Use for Value Identification:** A sophisticated model aims not just to predict the outcome, but to identify discrepancies between its own forecast and the market's forecast. This is the concept of "value betting".[37] By creating features that represent the difference between the model's predicted probability and the market's implied probability (e.g., model_prob_H - market_prob_H), the model can be trained to explicitly identify profitable opportunities.

By constructing a feature set that incorporates these four distinct pillars—Form, Process, Strength, and Market—the model receives a holistic and robust view of each matchup. It understands a team's recent results (Form), the quality of the process that generated those results (xG), its long-term historical standing (Elo), and the collective wisdom of the betting market (Odds). This multi-faceted approach prevents over-reliance on any single type of information and is a hallmark of state-of-the-art sports prediction systems.

The following table provides a blueprint for constructing such a feature set.

**Table 2: Compendium of Predictive Features**

| Feature Name | Category | Definition/Calculation Logic | Rationale for Inclusion |
|---|---|---|---|
| **Home_PPG_L5** | Form | Average points per game for the home team in their last 5 home matches. | Captures short-term home form and momentum. |
| **Away_GD_L10** | Form | Average goal difference per game for the away team in their last 10 away matches. | Measures recent performance balance over a medium-term window. |
| **Home_xGF_L10** | Advanced Metric | Average Expected Goals For per game for the home team in their | Measures the quality of scoring chances created, independent |

| | | | |
|---|---|---|---|
| | | last 10 home matches. | of finishing luck.[29] |
| **Away_xGA_L10** | Advanced Metric | Average Expected Goals Against per game for the away team in their last 10 away matches. | Measures the quality of chances conceded, reflecting defensive solidity.[32] |
| **Home_xGD_L10** | Advanced Metric | Home_xGF_L10 - Home_xGA_L10 | A powerful indicator of underlying performance dominance. |
| **Elo_Home** | Strength | The home team's Elo rating before the match. | Provides a dynamic, long-term measure of overall team strength.[7] |
| **Elo_Away** | Strength | The away team's Elo rating before the match. | Complements the home team's rating. |
| **Elo_Diff** | Strength | Elo_Home - Elo_Away | Directly quantifies the difference in strength between the two teams. |
| **Market_Prob_H** | Market | Implied probability of a home win, derived from closing betting odds. | Captures the market's sophisticated forecast, including non-statistical information.[21] |
| **Market_Prob_D** | Market | Implied probability of a draw, derived from closing betting odds. | Provides the market's assessment of the likelihood of a stalemate. |
| **Market_Prob_A** | Market | Implied probability of an away win, derived from closing betting odds. | Complements the home and draw market probabilities. |
| **SquadValue_Diff** | Strength | Difference in total market value between the home and away squads.[20] | A proxy for the talent and financial resources available to each team. |

# Part II: A Critical Review of Predictive Modeling Paradigms

Having established a robust data and feature engineering framework, the next step is to select and understand the predictive models that will consume this information. The field of football prediction is rich with different modeling paradigms, each with its own theoretical underpinnings, strengths, and weaknesses. This section provides a critical survey of the three main families of models: foundational statistical models, powerful machine learning classifiers, and the cutting-edge deep learning architectures.

## Section 4: The Canon of Statistical Goal-Based Models

This family of models represents the classical approach to football prediction. They do not attempt to classify the match outcome directly but instead model the underlying data-generating process of the game: the scoring of goals. By treating goal scoring as a stochastic process, these models provide highly interpretable parameters and form the theoretical basis for many modern hybrid systems. They can be thought of as creating a "physics engine" for the game, where the outputs are governed by parameters representing tangible concepts like "attack strength" and "defense strength." This interpretability is a key advantage over more opaque "black-box" machine learning methods.

### 4.1. The Independent Poisson Model

The Independent Poisson model is the genesis of statistical football modeling. Its core idea is elegant in its simplicity: the number of goals scored by the home team, $X$, and the away team, $Y$, are modeled as draws from two independent Poisson distributions.[5]

- Mathematical Formulation:
  $X \sim \text{Poisson}(\lambda h)$
  $Y \sim \text{Poisson}(\lambda a)$
  where $\lambda h$ and $\lambda a$ are the expected number of goals for the home and away teams, respectively.
- **Parameterization:** The power of the model comes from how these goal expectancies, $\lambda$, are estimated. Following the seminal work of Maher (1982), they are typically modeled using a regression framework where the log of the expectancy is a linear combination of team-specific parameters and a home-advantage term.[4] For a match between team $i$ (home) and team $j$ (away), the model is:
  $\log(\lambda h)=\text{home}+\text{atti}+\text{defj}$
  $\log(\lambda a)=\text{attj}+\text{defi}$
  Here, home is a constant parameter representing the home-field advantage, att represents a team's attacking strength, and def represents its defensive strength (often modeled as a negative value). These parameters are estimated from historical match

data using maximum likelihood estimation.

- **Strengths and Weaknesses:** The model's primary strengths are its simplicity and high interpretability.[3] It provides a solid theoretical baseline. However, it suffers from two key weaknesses: (1) it assumes the number of goals scored by each team is independent, which may not hold true in reality (e.g., a red card could affect both teams' scoring rates), and (2) it consistently underestimates the frequency of low-scoring draws, such as 0-0 and 1-1, which occur more often in real football matches than a standard Poisson process would predict.[3]

## 4.2. The Dixon and Coles Model

The model developed by Mark Dixon and Stuart Coles in their 1997 paper, "Modelling Association Football Scores and Inefficiencies in the Football Betting Market," is arguably the most influential and widely cited statistical model in the football analytics literature.[23] It directly addresses and corrects the main flaws of the Independent Poisson model, making it a much more robust and accurate tool.[3]

- **Core Improvements:**
    1. **Low-Score Adjustment:** The key innovation of the Dixon-Coles model is the introduction of a dependence parameter, ρ, which adjusts the probabilities of a few specific low-scoring outcomes. It recognizes that scores like 0-0, 1-0, 0-1, and 1-1 do not behave according to the independence assumption. The model "skims" probability from the independent Poisson calculation for these scores and redistributes it based on the value of ρ, bringing the model's predictions for these common results more in line with observed frequencies.[3]
    2. **Time-Decay Factor:** The model also introduces a weighting parameter, ξ (xi), that applies an exponential time decay to historical matches. This ensures that more recent games are given greater weight when estimating team strength parameters, while older games have less influence. This elegantly captures the dynamic and ever-changing nature of team form, a crucial element that the basic Poisson model ignores.[3]
- **Significance:** Due to these enhancements, the Dixon-Coles model provides a significantly better fit to real-world football data and is considered the benchmark against which new statistical models are often compared. Implementations and detailed discussions are available in various academic papers and online tutorials.[39]

## 4.3. The Bivariate Poisson Model

While the Dixon-Coles model adjusts for dependence in a specific, targeted way, the Bivariate Poisson model offers a more general approach. It directly models the correlation between the home and away goal counts by introducing a covariance term, λ3, into the joint probability distribution function.[3]

- Mathematical Formulation: The joint probability of a scoreline (x,y) is given by: P(X=x,Y=y)=e−(λ1+λ2+λ3)x!λ1xy!λ2yΣi=0min(x,y)(ix)(iy)i!(λ1λ2λ3)i
  Here, λ1 and λ2 relate to the marginal goal expectancies, and λ3 captures the covariance. A positive λ3 implies that if one team scores more, the other is also likely to score more, modeling an open, attacking game.
- **Application:** This model is particularly useful when a positive correlation between goal-scoring events is expected. It has been successfully integrated into modern hybrid frameworks where the statistical model's output is used as a feature for a machine learning classifier.[8] While more flexible in modeling dependence, it can also be more computationally intensive to fit than the Dixon-Coles model.

## Section 5: The Power of Machine Learning Classifiers

This paradigm shifts away from modeling the goal-scoring process and instead treats football prediction as a standard machine learning task. Given a rich, well-engineered feature set (as described in Section 3), these models learn complex, non-linear relationships between the input features and the match outcome. They are generally more powerful and flexible than purely statistical models, though often at the cost of some interpretability.

### 5.1. Ensemble Methods: The Workhorses of Tabular Data

Ensemble methods, which combine the predictions of multiple individual models, are the state-of-the-art for most problems involving structured, tabular data.

- **Random Forest (RF):** A Random Forest is an ensemble of many individual decision trees.[47] It operates by building a multitude of trees on bootstrapped samples of the training data and using a random subset of features for the splits at each node. The final prediction is made by aggregating the votes from all trees (for classification) or averaging their outputs (for regression). This process makes the model highly robust to overfitting, adept at capturing complex non-linear interactions, and capable of providing feature importance scores, which offer a degree of interpretability.[49] Numerous studies have demonstrated the high performance of RF models in football prediction, with some reporting accuracies upwards of 56% to 80%, depending on the context and dataset.[8]
- **Gradient Boosting Machines (GBM):** This family of algorithms (including popular implementations like XGBoost, LightGBM, and CatBoost) represents the pinnacle of performance for many tabular data tasks. Unlike Random Forests, which build trees in parallel, Gradient Boosting is an iterative process. It builds trees sequentially, where each new tree is trained to correct the errors made by the ensemble of preceding trees.[49] This focus on correcting residual errors allows boosting models to achieve

extremely high predictive accuracy. They are frequently used in football analytics and are often the top-performing models in academic comparisons and practical applications.[33]

### 5.2. Support Vector Machines (SVM)

Support Vector Machines are powerful classification algorithms that work by finding an optimal hyperplane that best separates data points of different classes in a high-dimensional space.[52]

- **Application in Football:** SVMs have been applied to football prediction with mixed results. Some studies have reported very high accuracies, in some cases over 70% [54], while others have found more modest performance.[53] The effectiveness of an SVM is critically dependent on the choice of the kernel function (e.g., linear, polynomial, radial basis function), which transforms the data into a higher-dimensional space where it might be linearly separable.[52] As SVMs are inherently binary classifiers, applying them to the three-way football outcome requires using strategies like one-vs-one or one-vs-rest, where multiple binary classifiers are trained to solve the multi-class problem.[55]

### 5.3. The Critical Step of Model Calibration

A common pitfall when using powerful machine learning models like Random Forests or Gradient Boosting is that while they may be excellent at classifying outcomes, their raw probability outputs are often poorly calibrated.[32] A model might correctly predict the winner 70% of the time it assigns a 70% probability, but it might also predict the winner 70% of the time it assigns a 90% probability. For applications in betting, where the precise probability is essential for identifying value, this is a fatal flaw.

- **The Solution: Calibration Techniques:** This issue can be rectified by a post-processing step known as calibration. This involves training a secondary, simpler model on a validation set to map the uncalibrated probabilities from the primary model to calibrated ones. The two most common methods are:
  1. **Platt Scaling:** Fits a logistic regression (sigmoid) function to the model's outputs.[32]
  2. **Isotonic Regression:** A non-parametric approach that fits a non-decreasing function.
- **Importance:** This step is crucial for transforming a good classifier into a reliable probabilistic forecasting tool. It is an essential component of any serious modeling pipeline aimed at generating probabilities for decision-making or betting.

The breadth of research reveals an important truth: there is no single "best" algorithm that dominates all others in every context. This is an expression of the "No Free Lunch" theorem in machine learning. Published studies report a wide range of top-performing models and accuracies: one paper finds Random Forest to be superior with 56.25% accuracy [8], another favors LightGBM at 52.8% [51], and a third suggests SVM can reach 71%.[54] This variation arises from differences in leagues, feature sets, validation periods, and evaluation metrics. Furthermore, simpler statistical models can sometimes outperform more complex machine learning models, especially on smaller datasets where the latter are prone to overfitting.[56] This leads to an inescapable conclusion: it is not enough to simply select a model based on its reported performance in the literature. It is imperative to implement a rigorous, project-specific validation framework to empirically determine which algorithm performs best on the unique data and feature set developed for the 3. Liga. The process of validation is more critical than the initial choice of algorithm.

## Section 6: The Frontier of Deep Learning Architectures

Deep learning represents the cutting edge of predictive modeling, offering the potential to learn hierarchical representations and complex patterns directly from data. While these models can be immensely powerful, they introduce significant complexity and are best considered a potential future step for this project rather than an initial approach. The marginal performance gains over a well-tuned Gradient Boosting model may not justify the substantial increase in data requirements, computational cost, and implementation effort, especially for a single-person project on a lower-tier league. Indeed, some studies show that state-of-the-art ensemble models like CatBoost can outperform deep learning architectures on football prediction tasks.[57]

### Neural Networks for Tabular Data

- **Multi-Layer Perceptrons (MLPs):** The foundational deep learning model is the MLP, a feed-forward neural network with one or more hidden layers. MLPs are universal approximators, meaning they can model any continuous function given enough complexity. They have been applied to tabular feature sets for football prediction with some success, offering a flexible way to capture non-linear relationships.[1]
- **Specialized Tabular Architectures:** In recent years, novel architectures have been developed specifically for tabular data, aiming to combine the power of deep learning with the structure of relational data. Models like **TabNet** [57] and tabular **Transformers** [60] use attention mechanisms to learn which features are most important for a given prediction and to model complex interactions between them automatically.

### Sequence Models for Time-Series Analysis

- **Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs):** These models are explicitly designed to handle sequential data. A football season can be framed as a time series, where each match is a point in the sequence. LSTMs can be used to model the evolution of team form over time, theoretically capturing long-term dependencies that fixed-window rolling averages might miss.[1] However, they are notoriously difficult to train and, for this type of problem, may not offer a significant advantage over a robust feature set that already includes dynamic metrics like Elo and time-decayed rolling averages.

## Hybrid Deep Learning Models

The true state-of-the-art in deep learning research often involves hybrid architectures. For instance, a model might use a one-dimensional Convolutional Neural Network (1D CNN) to extract local patterns and short-term dependencies from a sequence of match data, while a Transformer layer uses self-attention to model the long-range, global context across the season.[60] These models are highly complex but offer a powerful way to integrate different types of feature interactions.

The following table provides a comparative analysis of the different modeling paradigms, offering a strategic map of the landscape to guide model selection.

**Table 3: Comparative Analysis of Predictive Models**

| Model Paradigm | Specific Model | Core Idea | Key Assumptions | Primary Strengths | Primary Weaknesses | Typical Use Case |
|---|---|---|---|---|---|---|
| **Statistical** | Dixon & Coles | Models goal counts as a time-weighted, dependence-adjusted Poisson process. | Goals follow a Poisson distribution; strengths evolve over time. | High interpretability; strong theoretical basis; good baseline. | Less flexible than ML; may not capture all non-linearities. | Generating baseline goal expectancies; building interpretable models. |
| **ML Ensemble** | Random Forest | Averages predictions from many decorrelated decision trees. | Relationships can be captured by tree-based rules. | High accuracy; robust to overfitting; handles non-linear data well. | Less interpretable than statistical models; raw probabilities are uncalibrated. | Primary classifier for tabular data with a strong feature set. |
| **ML Ensemble** | Gradient Boosting | Sequentially builds | Errors of a weak learner | State-of-the-art | Prone to overfitting | The go-to choice for |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | models to correct prior errors. | ensemble can be reduced iteratively. | accuracy on most tabular tasks; highly flexible. | without careful tuning; can be a "black box." | maximizing predictive accuracy in a competition setting. |
| **Deep Learning** | MLP / Transformer | Learns hierarchical feature representations through a network of connected nodes. | Complex patterns are discoverable through non-linear transformations. | Can model extremely complex relationships; learns features automatically. | Data-hungry; computationally expensive; complex to implement and tune. | Frontier research; problems with massive datasets or image/sequence data. |

# Part III: A Proposed State-of-the-Art Hybrid System

This section synthesizes the analysis from the preceding parts into a concrete, expert-recommended system. The proposed architecture is a multi-stage hybrid model designed to leverage the respective strengths of both statistical and machine learning paradigms. The goal is not only to achieve high predictive accuracy but also to build a system capable of identifying betting value by comparing its own forecasts to those of the market.

## Section 7: An Expert's Approach: Architecting a Hybrid Predictive Engine

The recommended approach is a two-stage hybrid engine. The first stage uses a statistical model to generate a theory-driven forecast of goal expectancies. The second stage uses a powerful machine learning model that takes these expectancies as a key input, alongside other features, to produce the final, nuanced outcome probability.

### 7.1. Stage 1: The Statistical Foundation - Goal Expectancy Model

The foundation of the system will be a robust statistical model that provides a structured, interpretable estimate of the fundamental offensive and defensive capabilities of each team.

- **Model Choice: Dixon and Coles Model:** The Dixon and Coles model is selected as the statistical engine.[23] It is chosen over a simpler Independent Poisson model because its two key features—the exponential time-weighting of matches and the adjustment for low-scoring draws—are critical for accurately modeling the dynamic nature of a competitive football league.[3] This model provides a superior fit to real-world data compared to its predecessors.

- **Implementation and Outputs:** The model will be fitted to the entire historical dataset of 3. Liga matches. For any given future match, the primary outputs of this stage will be the expected number of goals for the home team (λh) and the away team (λa). These two values are not the final prediction; rather, they are high-level, distilled features that encapsulate each team's current attacking and defensive strength, their form (via the time-weighting), and the context of home advantage.
- **Generated Features for Stage 2:** The outputs from the Dixon-Coles model will be passed as features to the next stage. These include:
  1. DC_lambda_home: The expected goals for the home team.
  2. DC_lambda_away: The expected goals for the away team.
  3. DC_prob_H, DC_prob_D, DC_prob_A: The probabilities for a home win, draw, and away win, calculated by summing the probabilities of all corresponding scorelines from the full joint probability matrix generated by the model.[5]

### 7.2. Stage 2: The Machine Learning Core - Value Identification Model

The second stage employs a powerful machine learning classifier to synthesize all available information and produce the final prediction.

- **Model Choice: Gradient Boosting Machine (GBM):** A Gradient Boosting Machine, specifically an implementation like LightGBM or XGBoost, is the recommended choice for the core classifier. These models consistently deliver state-of-the-art performance on structured, tabular data. They are computationally efficient, highly flexible, and adept at handling a mix of numerical and categorical features without extensive pre-processing.[49]
- **Comprehensive Feature Set:** The GBM will be trained on the full, rich feature set developed in Section 3. This is the key to the hybrid approach's power. The model will see:
  1. **Foundational Features:** All form and context features (rolling averages of goals, points, etc., separated by home/away).
  2. **Advanced Metrics:** All features derived from Expected Goals (xG) and the Elo rating system.
  3. **Market Intelligence:** The implied probabilities derived from closing betting odds.
  4. **Statistical Model Outputs:** Crucially, the five features generated by the Stage 1 Dixon and Coles model. This integration of a statistical model's output as an input to a machine learning model is the essence of this powerful hybrid technique.[8]
- **Prediction Target:** The GBM will be configured as a multi-class classifier, trained to predict the final three-way match outcome: Home Win, Draw, or Away Win. Its output

will be a set of probabilities for each of these three classes.

### 7.3. The Synergy of the Hybrid Approach

This two-stage architecture creates a powerful synergy. The Dixon-Coles model provides a robust, theory-driven "best guess" based on the fundamental physics of goal scoring. The Gradient Boosting model then acts as a meta-learner, taking this statistical forecast and learning the complex, non-linear ways in which it should be adjusted based on other information.

For example, the GBM can learn patterns that the statistical model alone cannot. It might discover that when the Dixon-Coles model predicts a tight match (e.g., λh≈λa), but the market odds heavily favor the home team, the home team tends to underperform. This allows the model to learn the conditions under which the statistical model or the market is likely to be biased. This ability to model the *relationship between different forecasts* (its own statistical forecast and the market's forecast) is what enables the system to move beyond simple prediction and into the realm of identifying "value." It directly addresses the advanced concept of decorrelating a model's predictions from the bookmaker's to find profitable niches.[37]

## Section 8: Rigorous Model Validation and Performance Evaluation

A predictive model is only as credible as the rigor of its evaluation. It is essential to test the system in a way that simulates real-world performance and to use metrics that accurately reflect the model's predictive quality and potential profitability. A model can easily achieve high accuracy on data it has already seen, so out-of-sample testing is paramount.

### 8.1. Backtesting Methodology: Simulating the Future

Standard k-fold cross-validation, which randomly shuffles and splits the data, is inappropriate for sports prediction data. This is because the data has a natural temporal order, and using future matches to train a model that predicts past matches (a phenomenon known as data leakage) would lead to an overly optimistic and entirely invalid measure of performance.

- **Walk-Forward Validation:** The correct approach is a time-series-aware validation method, often called walk-forward validation or expanding window validation. The process is as follows:
    1. Select an initial period of data for training (e.g., seasons 2018-2019 through 2021-2022).
    2. Train the entire hybrid pipeline on this data.
    3. Test the trained model by making predictions on the next block of data (e.g., the

first half of the 2022-2023 season).
4. Expand the training set to include the test data from the previous step.
5. Retrain the model and test it on the next block of data (e.g., the second half of the 2022-2023 season).
6. This process is repeated, "walking forward" through time, ensuring the model is always trained only on information that would have been available at the time of prediction.

## 8.2. Choosing the Right Evaluation Metrics

The choice of metric should align with the model's purpose. Since the hybrid engine produces probabilities for ordered outcomes, metrics must be chosen that can properly evaluate the quality of these probabilistic forecasts.

- **Classification Accuracy:** This metric ((True Positives + True Negatives) / Total Samples) is simple and intuitive but can be highly misleading, especially if the classes are imbalanced (e.g., if home wins are much more common than draws).[22] It should be reported but never used as the sole measure of performance.
- **Probabilistic Scoring Rules:** These metrics evaluate the quality of the predicted probabilities themselves.
  - **Brier Score:** Measures the mean squared error between the predicted probabilities and the actual outcomes. It provides a comprehensive measure of calibration and refinement. Lower scores are better.
  - **Log-Loss:** A metric that heavily penalizes confident but incorrect predictions. It is a very sensitive measure of probabilistic accuracy.
  - **Ranked Probability Score (RPS):** This is the ideal metric for this specific problem. Football outcomes have a natural ordering (e.g., a home win is "further away" from an away win than a draw is). RPS is a scoring rule that accounts for this ranked nature. It penalizes a model more for being "very wrong" (e.g., predicting a strong home win when the away team wins) than for being "slightly wrong" (e.g., predicting a home win when the match is a draw).[32] A lower RPS indicates a more accurate and refined set of probabilities.
- **Calibration Plots:** These are visual tools used to assess whether a model's probability outputs are reliable. A calibration plot bins predictions by their predicted probability (e.g., all predictions between 20-30%) and plots this against the actual observed frequency of the event in those bins. In a perfectly calibrated model, the plot would form a straight diagonal line: when the model predicts a 30% probability, the event should actually occur 30% of the time.[31]

## 8.3. Simulating Financial Performance

Ultimately, a model designed with betting in mind must be evaluated on its ability to generate a positive financial return. This requires simulating a betting strategy based on the model's outputs.

- **Return on Investment (ROI):** The most direct measure of profitability. The simulation involves defining a "value threshold." For example, the strategy might be: "Place a bet on a home win if the model's predicted probability is at least 5% higher than the market's implied probability." The ROI is then calculated over the entire backtest period. Many studies use ROI as the ultimate benchmark of success.[1]
- **The Kelly Criterion:** Simply betting a flat stake on every "value" opportunity is a valid strategy, but it is not optimal. The Kelly Criterion is a formula from information theory that provides a sophisticated approach to bankroll management. It calculates the optimal fraction of one's bankroll to stake on a given bet, based on the perceived edge (the difference between the model's probability and the market's) and the odds offered. Using a fractional Kelly strategy (e.g., betting half of the amount suggested by the full Kelly formula) can maximize long-term growth while managing risk.[7] Simulating a Kelly-based strategy provides a more realistic assessment of the model's financial potential.

It is critical to recognize the duality of evaluation. A model must be assessed against two distinct but related criteria: predictive accuracy and profitability. A model can be highly accurate in its predictions but unprofitable if its forecasts are perfectly correlated with the betting market, offering no value edge.[37] Conversely, a model might have a low overall accuracy but be highly profitable if it can consistently identify a niche of mispriced outcomes (e.g., draws) that the market systematically underestimates.[9] Therefore, a comprehensive evaluation framework must include both probabilistic metrics like RPS to measure the intrinsic quality of the predictions and financial metrics like ROI to measure their practical utility against a real-world market.

The following table provides a glossary of the key evaluation metrics for this project.

**Table 4: Glossary of Model Evaluation Metrics**

| Metric Name | Type | Formula/Definition | Interpretation | Primary Use Case in this Project |
|---|---|---|---|---|
| **Accuracy** | Accuracy | (TP + TN) / (TP + TN + FP + FN) | Higher is better. | A simple, baseline understanding of classification performance. Not for primary model selection. |
| **Brier Score** | Probabilistic | Mean squared difference between predicted probabilities and | Lower is better (0 is perfect). | Measures overall quality of probabilistic forecasts, rewarding |

| | | | | |
|---|---|---|---|---|
| | | actual outcomes. | | calibration. |
| **Ranked Probability Score (RPS)** | Probabilistic | Measures the difference between cumulative probability forecasts and cumulative outcomes. | Lower is better (0 is perfect). | The primary metric for evaluating the model's predictive accuracy, as it accounts for the ordered nature of outcomes.[32] |
| **Calibration Plot** | Diagnostic | Visual plot of predicted probability vs. observed frequency. | A plot close to the diagonal y=x line indicates good calibration. | Visually assessing the reliability and trustworthiness of the model's probability outputs.[31] |
| **Return on Investment (ROI)** | Financial | (Total Profit / Total Stakes) * 100% | Higher is better. | The ultimate test of the model's ability to identify value and generate profit against market odds.[7] |

# Part IV: Conclusion and Future Horizons

## Section 9: Synthesis and Practical Recommendations

This report has outlined a comprehensive, expert-level framework for a personal project aimed at predicting match outcomes in Germany's 3. Liga. The proposed methodology is designed to be both academically rigorous and practically applicable, guiding the user from initial data acquisition through to the deployment and evaluation of a state-of-the-art hybrid predictive system.

The core of the recommended approach is a multi-faceted strategy that acknowledges the unique challenges of football analytics. The journey begins with confronting the low-scoring, high-variance nature of the sport, which necessitates a move beyond simple outcome data towards metrics that capture the underlying performance process. For a lower-tier competition like the 3. Liga, this requires a diligent and creative data engineering phase, fusing information from multiple sources—including web-scraped results, API-driven odds, and specialized data like squad market values—to compensate for the potential scarcity of advanced, pre-packaged statistics.

The most critical phase is feature engineering, where raw data is transformed into a rich set of predictive variables. Success hinges on creating a holistic feature set that captures four distinct aspects of team performance: recent **Form** (rolling averages), underlying **Process** quality (Expected Goals), long-term **Strength** (Elo ratings), and **Market** expectation (betting odds).

The recommended predictive engine is a two-stage hybrid model. **Stage 1** employs a **Dixon and Coles statistical model** to generate theory-driven goal expectancies and baseline probabilities. These outputs are then fed as powerful, high-level features into **Stage 2**, a **Gradient Boosting Machine**. This machine learning core learns the complex, non-linear relationships between the statistical forecast, market intelligence, and other situational features to produce a final, highly nuanced prediction. This hybrid architecture leverages the interpretability and theoretical grounding of statistical models with the raw predictive power of modern machine learning.

Finally, a rigorous evaluation framework is essential. This involves using a time-aware **walk-forward backtesting** methodology and employing appropriate metrics. Performance should be judged not on simple accuracy, but on the quality of the probabilistic forecasts using the **Ranked Probability Score (RPS)** and on financial viability through simulated **Return on Investment (ROI)**, potentially guided by the **Kelly Criterion**.

The actionable roadmap for this project is as follows:

1. **Data Infrastructure:** Systematically scrape and collate data from sources like FBref, TheSportsDB, and football-data.org. Focus on obtaining historical results, fixtures, and betting odds.
2. **Feature Engineering:** Develop a comprehensive feature set based on the four pillars (Form, Process, Strength, Market). Implement functions to calculate rolling averages, Elo ratings, and implied probabilities from odds. If possible, model a simplified xG metric.
3. **Stage 1 Modeling:** Implement and fit the Dixon and Coles model to the historical data to generate goal expectancy features.
4. **Stage 2 Modeling:** Train a Gradient Boosting classifier (e.g., LightGBM) using the full feature set, including the outputs from Stage 1.
5. **Calibration:** Apply Platt Scaling or Isotonic Regression to the GBM's probability outputs to ensure they are well-calibrated.
6. **Evaluation:** Conduct a rigorous walk-forward backtest, evaluating the final calibrated probabilities using RPS and simulating a betting strategy to calculate ROI.

The most critical determinants of success will be the diligence applied during the data engineering phase and the creativity and thoroughness of the feature engineering process. These foundational steps provide the high-quality fuel that a powerful model needs to perform effectively.

## Section 10: Beyond the Current Scope: Avenues for Further

## Research

Upon the successful implementation and validation of the core hybrid model, several exciting avenues for further research and enhancement become available. These advanced topics can significantly increase the model's granularity and predictive power.

- **Incorporating Player-Level Data:** The current model operates at the team level. A significant step forward would be to incorporate player-level data. This could include individual player ratings from sources like the FIFA video game series [8], market values from Transfermarkt [20], or detailed performance metrics from event data providers. A model that understands the quality of individual players can more accurately assess the impact of injuries, suspensions, or tactical selections involving key personnel. Hybrid models that integrate both team-level historical data and player-specific performance metrics have shown great promise in capturing nuanced interactions often overlooked by aggregate models.[61]
- **Complex Network Analysis:** A more abstract and powerful way to model team tactics is through complex network analysis. In this paradigm, a team's passing network can be represented as a graph, where players are the nodes and passes between them are the weighted edges. Metrics from network theory (e.g., centrality, clustering coefficient) can be calculated for these graphs to quantify tactical patterns, player importance, and team cohesion. Recent research has shown that combining these network metrics with traditional statistics can lead to more accurate predictive models, offering a deeper understanding of game patterns and strategies.[64]
- **Real-Time Dynamics and In-Play Modeling:** The current framework is designed for pre-match prediction. A fascinating extension is to build models that operate in real-time, updating predictions as a match unfolds. This would involve ingesting live data streams, such as in-game events (goals, cards, substitutions) or, more powerfully, the real-time movement of odds on a betting exchange. Models can be built to forecast odds movements themselves, treating the betting market as a financial time series and applying techniques from quantitative finance.[65] This opens the door to in-play betting strategies, which operate in a much more dynamic and data-rich environment.
- **Explainable AI (XAI):** While the Stage 1 statistical model is highly interpretable, the Stage 2 Gradient Boosting model is largely a "black box." Applying XAI techniques, such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), can provide valuable insights into the model's decision-making process. These tools can reveal *why* the model made a particular prediction for a specific match, highlighting which features were most influential. This not only builds confidence in the model but can also uncover novel strategic insights that are not immediately obvious from the data alone.[7]

By pursuing these advanced topics, the project can evolve from a powerful predictive system into a comprehensive football analytics platform, capable of generating deep, data-driven insights into the complexities of the beautiful game.

## Works cited

1. A Machine Learning Approach to Football Match Result Prediction - ResearchGate, accessed August 10, 2025, https://www.researchgate.net/publication/352940839_A_Machine_Learning_Approach_to_Football_Match_Result_Prediction
2. Predicting goal probabilities with improved xG models using event sequences in association football - PMC - PubMed Central, accessed August 10, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11524524/
3. Football Prediction Models: Which Ones Work the Best? | penaltyblog, accessed August 10, 2025, https://pena.lt/y/2025/03/10/which-model-should-you-use-to-predict-football-matches/
4. Predictive Modeling of Association Football Scores Using Bivariate Poisson, accessed August 10, 2025, http://article.sapub.org/10.5923.j.ajms.20201003.01.html
5. POISSON DISTRIBUTION IN FOOTBALL BETTING PREDICTIONS, accessed August 10, 2025, https://jns.edu.al/wp-content/uploads/2023/08/9_M_Shevroja_9535cd4b82.pdf
6. Predicting Football Match Results Using a Poisson Regression Model, accessed August 10, 2025, https://www.mdpi.com/2076-3417/14/16/7230
7. (PDF) Predicting Football Match Outcomes with eXplainable ..., accessed August 10, 2025, https://www.researchgate.net/publication/365850113_Predicting_Football_Match_Outcomes_with_eXplainable_Machine_Learning_and_the_Kelly_Index
8. (PDF) A Hybrid Machine Learning Framework for Soccer Match ..., accessed August 10, 2025, https://www.researchgate.net/publication/388315236_A_Hybrid_Machine_Learning_Framework_for_Soccer_Match_Outcome_Prediction_Incorporating_Bivariate_Poisson_Distribution
9. Machine learning model finds edge in draw markets (soccer), real or not ? : r/algobetting, accessed August 10, 2025, https://www.reddit.com/r/algobetting/comments/1mizt40/machine_learning_model_finds_edge_in_draw_markets/
10. 3. Liga Stats | FBref.com, accessed August 10, 2025, https://fbref.com/en/comps/59/3-Liga-Stats
11. Germany Liga 3 - Free API - TheSportsDB.com, accessed August 10, 2025, https://www.thesportsdb.com/league/4639-germany-liga-3
12. 3. Liga 2025/2026 Results Archive - Soccer24.com, accessed August 10, 2025, https://www.soccer24.com/germany/3-liga/archive/
13. Summary - 3. Liga - Germany - Results, fixtures, tables - Soccerway, accessed August 10, 2025, https://my.soccerway.com/national/germany/3-liga/20242025/regular-season/r818

41/

14. 3. Liga 2025/2026 live scores, results, Football Germany - Flashscore.com, accessed August 10, 2025, https://www.flashscore.com/football/germany/3-liga/
15. API Reference - football-data.org, accessed August 10, 2025, https://www.football-data.org/documentation/api
16. 3. Liga Football Data - Sportmonks' Football APIs, accessed August 10, 2025, https://www.sportmonks.com/football-api/3-liga-api/
17. Germany :: Football API / Livescore API, accessed August 10, 2025, https://live-score-api.com/leagues/league/1/Germany
18. Football/Soccer | Bundesliga Player Database - Kaggle, accessed August 10, 2025, https://www.kaggle.com/datasets/oles04/bundesliga-soccer-player
19. German Football Scores - Kaggle, accessed August 10, 2025, https://www.kaggle.com/dhruvildave/german-football-scores/tasks
20. 3. Liga 25/26 - Transfermarkt, accessed August 10, 2025, https://www.transfermarkt.us/3-liga/startseite/wettbewerb/L3
21. A Systematic Review of Machine Learning in Soccer Betting ..., accessed August 10, 2025, https://www.preprints.org/manuscript/202501.2060/v1
22. How to Use Python and Machine Learning to Predict Football Match ..., accessed August 10, 2025, https://www.kdnuggets.com/2023/01/python-machine-learning-predict-football-match-winners.html
23. Dixon-Coles Model explain by TOP LLM models (ChatGPT, Claude 3.5 Sonnet, Nemotron, Gemini Pro etc) | by Cristian Nedelcu | Medium, accessed August 10, 2025, https://medium.com/@cristian.nedelcu/dixon-coles-model-explain-by-top-llm-models-chatgpt-claude-3-5-sonnet-nemotron-gemini-pro-etc-2b264db5a28b
24. Is Football Unpredictable? Predicting Matches Using Neural Networks - MDPI, accessed August 10, 2025, https://www.mdpi.com/2571-9394/6/4/57
25. How we calculate Expected Goals (xG) - Fantasy Football Fix, accessed August 10, 2025, https://www.fantasyfootballfix.com/blog-index/how-we-calculate-expected-goals-xg/
26. What are Expected Goals (xG)? - Hudl, accessed August 10, 2025, https://www.hudl.com/blog/expected-goals-xg-explained
27. www.hudl.com, accessed August 10, 2025, https://www.hudl.com/blog/expected-goals-xg-explained#:~:text=An%20xG%20model%20uses%20historical,twice%20in%20every%2010%20attempts.
28. Expected goals - Wikipedia, accessed August 10, 2025, https://en.wikipedia.org/wiki/Expected_goals
29. What are Expected Goals (xG)? - use xG to predict match outcomes, accessed August 10, 2025, https://footballxg.com/what_are_expected_goals/
30. Does xG have a predictive value? : r/PremierLeague - Reddit, accessed August 10, 2025, https://www.reddit.com/r/PremierLeague/comments/1im5o2v/does_xg_have_a_predictive_value/

31. Building a Football xG Model with XGBoost | by Philiprj - Medium, accessed August 10, 2025, https://medium.com/@philiprj2/building-a-football-xg-model-with-xgboost-5939161bb655
32. Predicting football matches with Random Forests and Platt's scaling ..., accessed August 10, 2025, https://andrewwoods1.github.io/WIP_Predicting_RF/
33. Predicting Football Match Outcomes Using Event Data and Machine Learning Algorithms - Ulster University, accessed August 10, 2025, https://pure.ulster.ac.uk/files/213544028/Predicting_football_match_outcomes.pdf
34. Master Football Betting Predictions With These Expert Tips - YouTube, accessed August 10, 2025, https://www.youtube.com/watch?v=BgaTqw74UQE&pp=0gcJCfwAo7VqN5tD
35. How does machine learning work in sports betting? - OddsMatrix, accessed August 10, 2025, https://oddsmatrix.com/machine-learning-sports-betting/
36. A Systematic Review of Machine Learning in Sports Betting: Techniques, Challenges, and Future Directions - arXiv, accessed August 10, 2025, https://arxiv.org/html/2410.21484v1
37. (PDF) Exploiting sports-betting market using machine learning - ResearchGate, accessed August 10, 2025, https://www.researchgate.net/publication/331218530_Exploiting_sports-betting_market_using_machine_learning
38. Modelling Association Football Scores and Inefficiencies in the Football Betting Market (1996) | Mark J. Dixon | 512 Citations - SciSpace, accessed August 10, 2025, https://scispace.com/papers/modelling-association-football-scores-and-inefficiencies-in-58z5yfzu94
39. Predicting Football Results Using Python and the Dixon and Coles Model | penaltyblog, accessed August 10, 2025, https://pena.lt/y/2021/06/24/predicting-football-results-using-python-and-dixon-and-coles/
40. Extending the Dixon and Coles model: an application to women's football data - arXiv, accessed August 10, 2025, https://arxiv.org/pdf/2307.02139
41. Modelling Association Football Scores and Inefficiencies in the Football Betting Market., accessed August 10, 2025, https://www.ajbuckeconbikesail.net/wkpapers/Airports/MVPoisson/soccer_betting.pdf
42. Modelling Association Football Scores and Inefficiencies in the Football Betting Market - Oxford Academic, accessed August 10, 2025, https://academic.oup.com/jrsssc/article-pdf/46/2/265/48750190/jrsssc_46_2_265.pdf
43. Predicting and Modelling Football Matches with the R \textrm{R} Package footBayes, accessed August 10, 2025, https://www.researchgate.net/publication/392720368_Predicting_and_Modelling_Football_Matches_with_the_textrmR_Package_footBayes

44. Modelling Association Football Scores and Inefficiencies in the Football Betting Market | Journal of the Royal Statistical Society Series C - Oxford Academic, accessed August 10, 2025, https://academic.oup.com/jrsssc/article/46/2/265/6990546
45. Modelling Association Football Scores and Inefficiencies in The Football Betting Market, accessed August 10, 2025, https://fr.scribd.com/doc/270732164/Modelling-Association-Football-Scores-and-Inefficiencies-in-the-Football-betting-market
46. Dixon-Coles model | opisthokonta.net, accessed August 10, 2025, https://opisthokonta.net/?cat=48
47. Predicting Football Games: A Comparison Between LSTM and Random Forest - DiVA portal, accessed August 10, 2025, https://su.diva-portal.org/smash/get/diva2:1969719/FULLTEXT01.pdf
48. Using random forests to estimate win probability before each play of an NFL game - Iowa State University Digital Repository, accessed August 10, 2025, https://dr.lib.iastate.edu/bitstreams/17e429ca-3ad5-481a-b06a-a19f872c6351/download
49. Soccer match outcome prediction with random forest and gradient boosting models, accessed August 10, 2025, https://www.researchgate.net/publication/378355415_Soccer_match_outcome_prediction_with_random_forest_and_gradient_boosting_models
50. (PDF) An Overview of Machine Learning Applications in the Football Field - ResearchGate, accessed August 10, 2025, https://www.researchgate.net/publication/372823286_An_Overview_of_Machine_Learning_Applications_in_the_Football_Field
51. Sports Betting: An Application of Machine Learning to the Game Prediction, accessed August 10, 2025, https://www.ewadirect.com/proceedings/ace/article/view/20626
52. Support Vector Machine – Based Prediction System for a Football Match Result, accessed August 10, 2025, https://www.semanticscholar.org/paper/Support-Vector-Machine-%E2%80%93-Based-Prediction-System-a-Igiri/757be76ed214321a57a68893b6c546dc9840ed2f
53. Support Vector Machine–Based Prediction System for a Football Match Result - IOSR Journal, accessed August 10, 2025, https://www.iosrjournals.org/iosr-jce/papers/Vol17-issue3/Version-3/D017332126.pdf
54. Prediction of Winning Team using Machine Learning - International Journal of Engineering Research & Technology, accessed August 10, 2025, https://www.ijert.org/research/prediction-of-winning-team-using-machine-learning-IJERTCONV9IS03096.pdf
55. Enhancing Football Match Outcome Prediction - DiVA, accessed August 10, 2025, https://kth.diva-portal.org/smash/get/diva2:1886691/FULLTEXT01.pdf
56. (PDF) Comparing the performance of machine learning and statistical models in predicting football games - ResearchGate, accessed August 10, 2025, https://www.researchgate.net/publication/377351331_Comparing_the_performan

ce_of_machine_learning_and_statistical_models_in_predicting_football_games

57. Machine Learning for Soccer Match Result Prediction - arXiv, accessed August 10, 2025, https://arxiv.org/pdf/2403.07669
58. Gridiron Genius: Using Neural Networks to Predict College Football - Deep Blue Repositories, accessed August 10, 2025, https://deepblue.lib.umich.edu/bitstream/handle/2027.42/176935/Luke_Boll_Honors_Capstone_Report_-_Luke_Boll.pdf?sequence=1
59. Predicting Outcomes of Football Matches - CS230 Deep Learning - Stanford University, accessed August 10, 2025, https://cs230.stanford.edu/projects_fall_2018/reports/12445633.pdf
60. Predicting sport event outcomes using deep learning - PeerJ, accessed August 10, 2025, https://peerj.com/articles/cs-3011.pdf
61. Hybrid Machine Learning Forecasts for the UEFA EURO 2020 - arXiv, accessed August 10, 2025, https://arxiv.org/pdf/2106.05799
62. Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics - MDPI, accessed August 10, 2025, https://www.mdpi.com/2076-3417/10/1/46
63. A Generalizable Machine Learning Approach for Match Outcome Prediction with Insights from the FIFA World Cup - arXiv, accessed August 10, 2025, https://arxiv.org/pdf/2505.01902
64. Predicting soccer matches with complex networks and machine learning - arXiv, accessed August 10, 2025, https://arxiv.org/abs/2409.13098
65. Modeling and Forecasting Odds Movements on a ... - DiVA portal, accessed August 10, 2025, http://www.diva-portal.org/smash/get/diva2:1961953/FULLTEXT01.pdf
66. Data-Driven Team Selection in Fantasy Premier League Using Integer Programming and Predictive Modeling Approach - arXiv, accessed August 10, 2025, https://arxiv.org/html/2505.02170v1