Accuracy Without Profit: A Statistical Evaluation of Machine Learning in the English Premier League Betting Market

Mostafa Shams[1]

[1]Faculty of Computer and Information Menoufia University, Egypt
mustafa.samy2022385@ci.menofia.edu.eg

Abstract

This study explores the practical limits of predictive modeling in environments that are actively trying to beat the predictor. We tested three standard machine learning algorithms (XGBoost, LightGBM, and Random Forest) across two very different datasets to understand where the line between signal and noise truly lies. First, we tested the models on a "Crash" game—a source of verified, pure random noise—to ensure our methods did not find patterns where none existed. Second, we applied these models to twenty years of English Premier League match data (2000–2021), a highly efficient market where bookmakers constantly adjust odds to remove profitable opportunities.

Using a chronological testing method that mimics real-world conditions, we found that the models are statistically distinct from the bookmakers; they are not simply copying the market odds. However, while the models achieved respectable accuracy in predicting match outcomes, they failed to generate a *commercially viable* profit in the long run.. Our analysis reveals two primary reasons for this. First, we observed a clear "decay" in performance: a strategy that would have been profitable between 2006 and 2014 stopped working after 2015, suggesting the market became more efficient over time. Second, we found a significant "calibration error" of approximately 11%. Simply put, the models were consistently overconfident—assigning high probabilities to events that happened less often than predicted. This overconfidence caused standard risk-management strategies (like the Kelly Criterion) to bet too aggressively and lose money. These findings suggest that in mature, modern markets, the challenge for data scientists is no longer just predicting the winner, but accurately measuring how confident the model should be.

Reproducibility: All code, datasets, and full statistical reports generated during this study are publicly available for review and replication at: https://github.com/MostafaShams5/Predicting-EPL-Winners-But-Losing-Money-With-Data-Science.

1 Introduction

Predicting the future is the central goal of data science. Whether it is forecasting stock prices, weather patterns, or customer behavior, the aim is always the same: to find a hidden signal inside a noisy dataset. However, not all datasets are created equal. Some are static, waiting to be solved. Others are "adversarial"—meaning there is an opponent on the other side actively trying to prevent you from predicting correctly.

Sports betting provides a perfect laboratory for studying these adversarial environments. Unlike the stock market, which acts as a continuous flow of prices, a sports match has a definite start and end. The "price" (the odds) is set by bookmakers who are highly motivated to be accurate. If a predictive model can consistently beat these odds, it proves that the market has a leak—an inefficiency that can be exploited. Conversely, if a sophisticated model cannot make a profit, it suggests the market has become "efficient," meaning the price already reflects all available information.

For years, researchers have asked: ”Can machine learning beat the sports betting market?” This paper asks a different, perhaps more important question: ”How has the market learned to beat machine learning?”

To answer this, we designed a rigorous experiment using two very different types of data. First, we tested our models on a ”Crash” game—a simple online gambling game known to be purely random. This served as our ”control group.” We needed to be certain that our models would not hallucinate patterns where none existed. If they found a ”strategy” in pure noise, we would know our methods were flawed.

Second, we applied the same models to twenty-one years of English Premier League football data (2000–2021). Our approach was strictly historical. We did not let the models ”peek” at the future. We trained them season by season, exactly as a data scientist would have done in real time.

The results tell a story of a closing window. We found that while complex algorithms are indeed distinct from the bookmaker's odds—meaning they have their own ”opinion”—they are no longer profitable. Our analysis reveals a clear timeline: strategies that generated reliable profits in the mid-2000s slowly degraded until the edge effectively evaporated around 2015. Furthermore, we identified a critical flaw in how these models ”think.” While they are often correct about who will win, they are consistently overconfident about their chances. This gap between confidence and reality—known as calibration error—is what ultimately causes financial strategies to fail in a modern, efficient market.

2 Related Work and Background

Before diving into our own experiments, it is necessary to contextualize our findings within the broader literature of sports economics and statistical modeling. While the application of modern Machine Learning to sports betting is a relatively new field, it rests on two well-established pillars of research: the economic theory of Efficient Markets and the statistical problem of Probability Calibration.

2.1 The Efficient Market and the Strategic Bookmaker

The primary obstacle to profitable prediction is the Efficient Market Hypothesis (EMH). Originally proposed for financial markets, the theory suggests that asset prices already contain all known information, making consistent profit impossible without inside information [2]. In the specific context of sports, Sauer (1998) refined this into the ”Efficient Forecast Hypothesis,” arguing that wagering markets are efficient if the closing odds are unbiased predictors of match outcomes [10].

However, sports betting markets differ from financial markets in one critical way: the presence of an active opponent. Levitt (2004) argues that bookmakers are not merely passive market-makers but ”adversarial agents.” They are highly skilled at setting prices that not only reflect true probabilities but also exploit human biases—such as the ”favorite-longshot bias”—to maximize their own profit while minimizing risk [11, 3]. This creates a strategic environment where a predictive model is not just solving a math problem, but battling an optimized pricing algorithm.

Early research often found small cracks in this armor. Studies from the 1990s and early 2000s identified persistent inefficiencies [4]. However, as our study aims to demonstrate, the gap between ”market efficiency” and ”model accuracy” has closed significantly in the big-data era.

2.2 From Statistical History to Calibration Errors

To beat these efficient markets, researchers initially relied on parametric statistical models. The foundational work by Dixon and Coles (1997) established that simple averages are insufficient; models must account for time-decay, weighting recent match results more heavily than older ones [12]. This concept of ”recent form” remains a standard feature in modern engineering, including our own.

As the field moved from statistical models to Machine Learning (e.g., Random Forests and Neural Networks), the metric for success shifted. In standard classification tasks, we optimize for Accuracy. However, recent literature argues that in betting environments, Accuracy is a misleading metric. Wheatcroft (2020) demonstrated that a strategy's profitability is far more sensitive to ”Calibration Error” than to raw predictive accuracy [13].

This distinction is critical. If a model predicts a 60% probability of a win, but the event occurs only 50% of the time, the model is ”miscalibrated” [5]. Walsh and Joshi (2024) recently confirmed that when using risk-management strategies like the Kelly Criterion [6], calibration-optimized models significantly outperform accuracy-optimized ones [14]. If a model overestimates its own certainty, the Kelly criterion will command it to bet too aggressively, transforming a potentially winning strategy into a losing one.

3 Methodology and Data

To ensure our results were solid, we needed to compare a world of pure noise against a world of potential signal. We also needed a testing method that strictly followed the flow of time, preventing the common mistake of using future data to predict the past.

3.1 The Control Group: A Test of Pure Noise

Before analyzing football, we put our algorithms through a ”sanity check.” We collected data via live web scraping from the ”Crash” game hosted on 1xBet, a high-volume international betting platform popular in the Middle East. In this game, a multiplier starts at 1.00x and increases until it ”crashes” at a random point. We collected 2,596 consecutive rounds of this game.

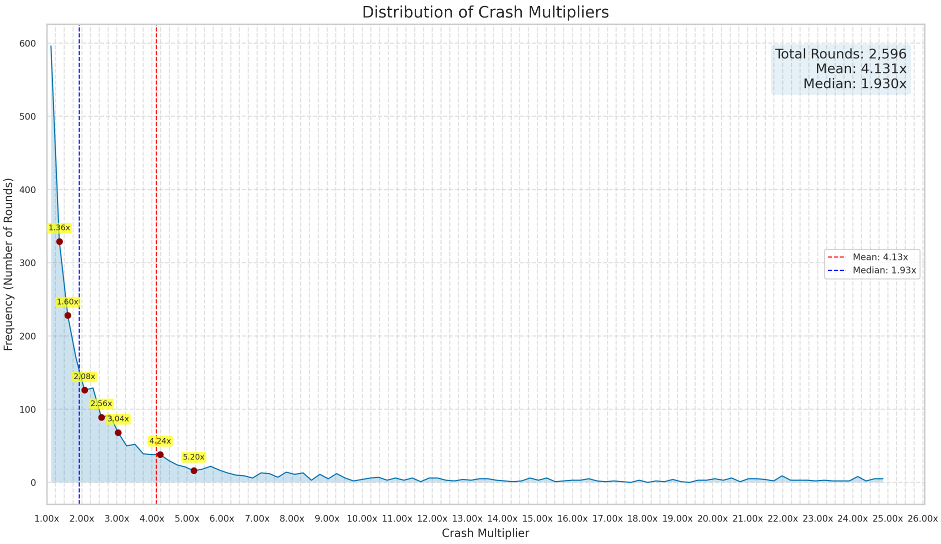


Figure 1: Distribution of Crash Game Multipliers. The game produces a massive number of low-value results (the median is 1.93x) with a long tail of high results that pull the average up. This purely random distribution served as our control group.

We used this control group to ensure our methods did not identify false patterns. As expected, the models failed to find any predictive edge in this random environment, validating that they were functioning correctly (see Appendix A).

3.2 The Real-World Data: English Premier League (2000–2021)

Our primary dataset covers twenty-one seasons of English top-flight football [8]. The original raw data contained detailed match statistics and odds from multiple bookmakers (including Bet365, William Hill, and VC Bet).

To create a consistent input for the models, we performed the following preprocessing steps:

1. **Market Consensus (Odds):** We utilized the average odds across all available major bookmakers for Home Win, Draw, and Away Win. This ensures the model is training on the global market opinion rather than a single provider's price.

2. **Team Fundamentals (Feature Engineering):** We engineered rolling window features representing each team's form over the Last 5 Games (L5). Specifically, we computed L5 averages for Points, Goals Conceded, Shots, Shots on Target, and Corners. This transforms raw match results into trend metrics.

Table 1: Snapshot of Transformed Input Data (First 10 Rows)

Date	**Home**	**Away**	**FTR**	**HO**	**DO**	**AO**	**H_Pts**	**H_GC**	**H_S**	**H_ST**	**H_C**	**A_Pts**	**A_GC**	**A_S**	**A_ST**	**A_C**
2001-08-20	Everton	Tottenham	D	2.25	3.20	2.75	3.0	1.0	12.0	9.0	4.0	1.0	0.0	8.0	2.0	6.0
2001-08-21	Ipswich	Derby	H	1.66	3.30	4.50	0.0	1.0	10.0	4.0	1.0	3.0	1.0	7.0	3.0	4.0
2001-08-21	Bolton	Middlesbr.	H	2.20	3.25	2.75	3.0	0.0	18.0	8.0	5.0	0.0	4.0	6.0	2.0	2.0
2001-08-21	Arsenal	Leeds	A	1.72	3.25	4.20	3.0	0.0	14.0	9.0	6.0	3.0	0.0	16.0	6.0	10.0
2001-08-22	Blackburn	Man Utd	D	4.50	3.30	1.66	0.0	2.0	14.0	4.0	10.0	3.0	2.0	12.0	6.0	6.0
2001-08-22	Fulham	Sunderland	H	1.90	3.20	3.50	0.0	3.0	8.0	5.0	4.0	3.0	0.0	12.0	6.0	1.0
2001-08-25	Arsenal	Leicester	H	1.22	5.00	9.50	1.5	1.0	14.0	7.0	8.0	0.0	5.0	6.0	1.0	3.0
2001-08-25	Blackburn	Tottenham	H	2.30	3.10	2.75	0.5	2.0	13.0	4.5	6.5	1.0	0.5	7.5	2.0	3.0
2001-08-25	Everton	Middlesbr.	H	2.00	3.10	3.30	2.0	1.0	15.0	10.0	6.0	0.0	2.5	8.5	3.0	2.0
2001-08-25	Fulham	Derby	D	1.50	3.40	6.00	1.5	1.5	10.5	5.0	6.0	1.5	2.0	4.5	2.5	3.0

Note: FTR = Full Time Result. HO/DO/AO = Home/Draw/Away Odds (Averaged). H_/A_ = Home/Away Team L5 Rolling Averages. Pts=Points, GC=Goals Conceded, S=Shots, ST=Shots on Target, C=Corners.

3.3 Model Specifications and Hyperparameters

To ensure reproducibility and isolate the effect of market efficiency from model complexity, we utilized standard, widely accessible machine learning architectures. All experiments were conducted with a fixed random seed (SEED = 42) for NumPy, TensorFlow, and library-specific generators.

Environment 1: The Crash Game (Time-Series)

For the crash game analysis, data was normalized to the range $[0, 1]$ using Min-Max scaling. The time-series data was transformed into sliding windows with a sequence length (T) of 15 rounds.

- **LSTM Regressor:** Designed to test for temporal dependencies. The architecture consists of an LSTM layer (64 units, return sequences=True), a Dropout layer (0.15), a second LSTM layer (32 units), a second Dropout layer (0.15), and a Dense layer (16 units, ReLU activation) feeding into a single linear output. It was optimized using Adam with Mean Squared Error (MSE) loss and Early Stopping (patience=4).

- **LSTM Autoencoder:** Designed to test for signal compressibility. The encoder uses an LSTM (64 units) to compress the sequence into a latent vector. This is fed into a RepeatVector (T) and passed to a decoder LSTM (64 units) followed by a TimeDistributed Dense layer. It was optimized using Adam with Mean Absolute Error (MAE) loss.

Environment 2: Premier League Betting (Classification)

We treated the sports betting problem as a multi-class classification task with three target classes: Home Win (H), Draw (D), and Away Win (A). The feature set consisted of 13 variables: 3 market-implied odds and 10 fundamental team metrics (rolling 5-game averages for points, goals, shots, and corners).

- **XGBoost:** Trained using the `multi:softprob` objective to output probabilities, using `mlogloss` as the evaluation metric.

- **LightGBM:** Trained using the `multiclass` objective with Gradient Boosting Decision Trees (GBDT).

- **Random Forest:** Constructed as a bagging ensemble of 100 trees using Gini Impurity for split quality.

In the sports betting experiment, we intentionally abstained from extensive hyperparameter grid-searching. The goal was to test whether the *information* contained in the features was sufficient for a standard algorithm to extract an edge, rather than testing the limits of hyperparameter optimization.

3.4 Validation Protocol: Preventing Look-Ahead Bias

The most common error in forecasting studies is "cheating" by shuffling the data. If you train a model on data from 2010 and 2020 mixed together to predict a game in 2015, the model implicitly learns the future trends.

To avoid this, we used a "Walk-Forward" approach:

- We started by training the model *only* on data from 2000 to 2005.

- We then asked it to predict the games in 2006.

- After 2006 finished, we added that data to the training set and asked it to predict 2007.

This process was repeated season by season until 2021. This mimics exactly how a real-world analyst operates: making decisions based only on what happened yesterday, never on what will happen tomorrow.

4 Results and Analysis

We analyzed the performance of three machine learning models—XGBoost, LightGBM, and Random Forest—over the 15-year validation period (2006–2021). The results reveal a disconnect between predictive accuracy and financial success.

As shown in Table 2, the model with the highest raw accuracy (Random Forest, 52.83%) generated the largest financial loss (-\$11,049). Conversely, XGBoost, which had the lowest accuracy, was the only model to generate a profit. This validates the "Accuracy Paradox" in betting markets: optimizing for the frequency of wins often leads to betting on "safe" favorites with poor value, resulting in negligible returns or a slow bleed of capital.

Table 2: Walk-Forward Validation Results (2006-2021)

Model	Accuracy	ROI (%)	Total PnL ($)	Brier Score
XGBoost	51.06%	**+0.29%**	**+$1,611**	0.638
LightGBM	51.79%	-1.10%	-$6,092	0.609
Random Forest	**52.83%**	-2.01%	-$11,049	**0.586**

4.1 Are the Models Just Copying the Odds?

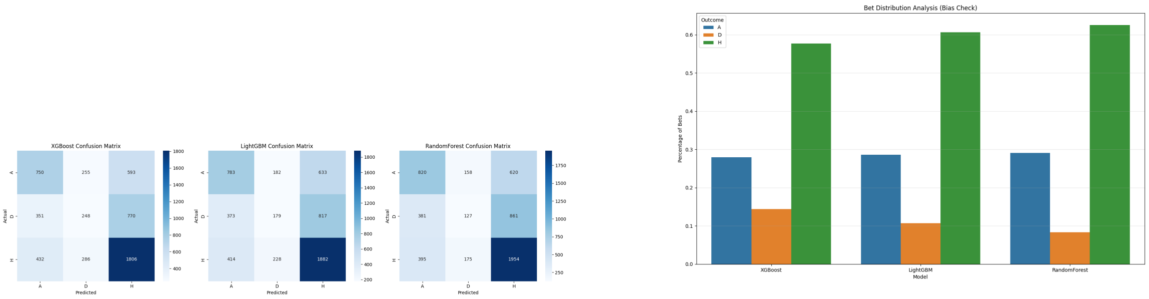
A primary concern was whether the models were simply mimicking the bookmaker's probabilities to minimize error. To test for informational independence, we utilized the Diebold-Mariano test [7].

The test returned significant statistics for both XGBoost (11.58, $p < 0.0001$) and LightGBM (8.74, $p < 0.0001$). Statistically, this confirms that the models are distinct from the market; they are identifying unique patterns in the Rolling Window (L5) features rather than just regressing to the implied odds.

However, distinctness does not imply superiority. When comparing the best model (XGBoost) against a common "Bet on Favorite" strategy, the difference yielded a p-value of 0.054. Strictly speaking, this falls just short of the 0.05 threshold for statistical significance. This suggests that while the AI found a signal, that signal was not strong enough to decisively outperform a simple heuristic over the full sample size.

4.2 The "Draw" Blind Spot

To understand why high accuracy failed to translate into robust profits, we analyzed the specific betting behaviors of the models (Fig. 2).



(a) Confusion Matrices

(b) Bet Distribution by Outcome

Figure 2: (a) The Confusion Matrices reveal a systemic failure to predict Draws (center column). (b) The Bet Distribution shows the models compensating for this uncertainty by over-betting on the Home Team (right bars), effectively "chasing" the home-field advantage.

The Confusion Matrices (Fig. 2a) reveal a flaw: the models are "risk-averse." They almost never successfully predict a Draw, which is typically the outcome with the highest variance but also significant value. Instead, as seen in Fig. 2b, they aggressively bet on the Home Team. By avoiding the risky Draw calls, the models artificially inflate their accuracy (since Home wins are common) but miss the high-odds opportunities necessary to overcome the bookmaker's margin.

4.3 The Timeline of Profitability (Alpha Decay)

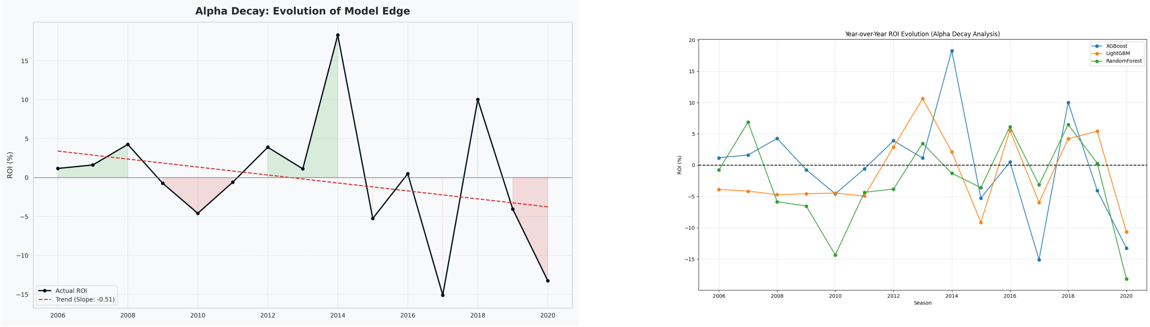
While the models were distinct, they were not consistently profitable. When we broke down the returns year by year, a certain pattern emerged (Fig. 3). This phenomenon is often called

”Alpha Decay”—the tendency for a profitable strategy to stop working as the market adapts.

- **Phase I: Market Inefficiency (2006–2014):** In the earlier years of our simulation, the XGBoost model showed strong performance. It generated positive returns in several seasons, peaking with an +18.31% return on investment in 2014. During this era, the ”edge” was real. Crucially, the model's advantage during this period was large enough to overcome the ”vigorish” (the bookmaker's built-in fee of approx. 4%–7%), resulting in net profit.
- **Phase II: Market Correction (2015–2021):** Around 2015, the trend reversed. The returns became highly volatile and failed to sustain growth, suffering significant drawdowns in 2017 (-15.10%) and 2020 (-13.24%). While the models still found patterns, the signal was too weak to overcome the bookmaker's margin.

This timeline aligns with the broader modernization of sports betting. As bookmakers integrated their own advanced algorithms and real-time data feeds around the mid-2010s, the inefficiencies that our models exploited in the early data simply vanished.

To verify that the models didn't simply degrade in quality, we analyzed the stability of their error rates. We found that the Mean Brier Error remained remarkably stable, moving only from 0.242 in the profitable early era to 0.234 in the losing late era. The models didn't change; the hurdles they had to jump just got higher.



(a) Long-term Trend

(b) Year-over-Year Volatility

Figure 3: Alpha Decay Analysis. (a) The stylized trend showing the ”Golden Age” (green) vs the ”Efficient Era” (red). (b) The raw performance lines for all three models, showing the specific volatility and the sharp drop-off for XGBoost (blue) after 2014.

We can see this impact most clearly when looking at the cumulative wealth over time (Fig. 4). If you had started using the XGBoost model in 2006, you would have seen a steady climb in profits. However, the data reveals a persistent ”Alpha Decay.” A linear regression of the year-over-year returns shows a negative slope of **-0.51**, indicating that the model's edge over the bookmaker degraded by approximately half a percentage point every single season as the market became more efficient.

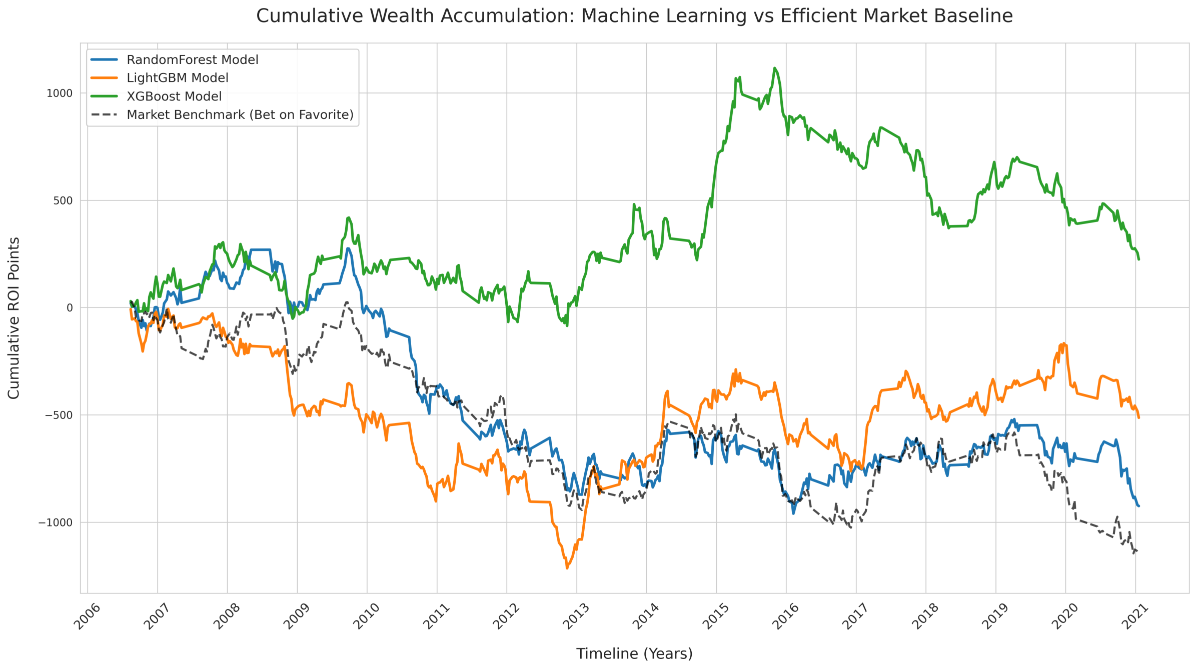


Figure 4: Cumulative Wealth Accumulation. The green line (XGBoost) illustrates the ”Golden Age” of sports betting, showing strong growth until 2015. After that point, the line flattens and becomes volatile, struggling to beat the simple ”Market Benchmark” (grey dashed line) of just betting on the favorite.

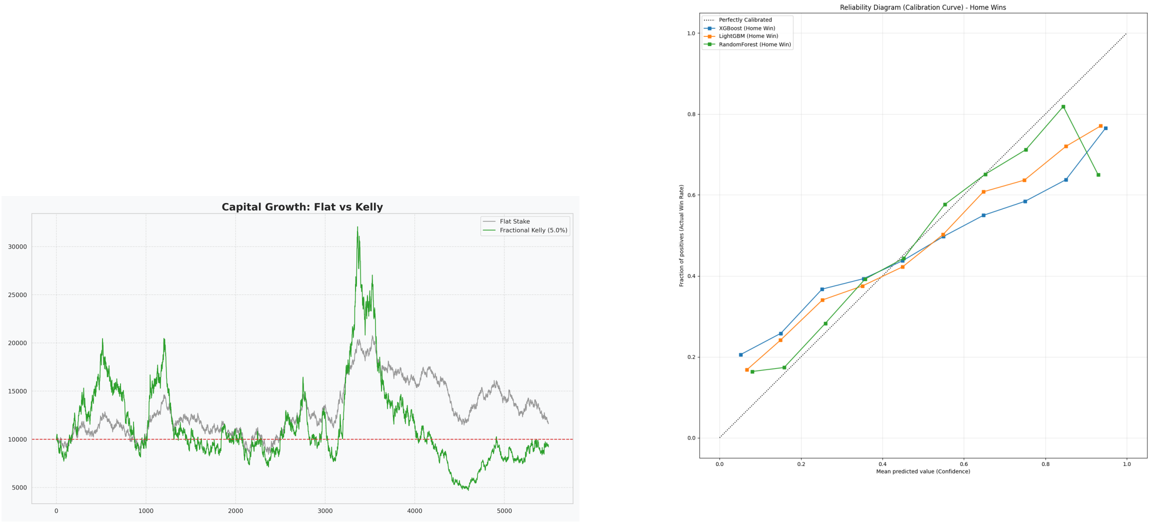
4.4 Impact of Calibration Error on ROI

The most critical finding of our study lies in the difference between being ”accurate” and being ”calibrated.” We simulated two betting strategies (Fig. 5a). The first was ”Flat Betting,” where the model bet the same amount ($100) on every game it liked. The second was the ”Kelly Criterion,” a famous mathematical formula that increases the bet size when the model is more confident. In theory, the Kelly strategy should make much more money.

In our experiment, the opposite happened:

- **Flat Betting Final Bankroll:** $11,611
- **Kelly Betting Final Bankroll:** $9,206

Why did the ”smarter” mathematical strategy lose $2,400 compared to the simple one? The answer is the ”Expected Calibration Error” (ECE), which we measured at **0.1109**. This means that when the model said, ”I am 60% sure this team will win,” the team actually won only 49% of the time. The model was chronically overconfident. Because it overestimated its own edge, the Kelly formula commanded it to make bets that were too large, amplifying the losses.



(a) Capital Growth: Flat vs. Kelly

(b) Calibration Curve

Figure 5: (a) The Kelly strategy (green line) underperformed the Flat strategy (grey line), a classic sign of overconfidence. (b) The calibration curve confirms this: the model's confidence lines are mostly below the "Perfectly Calibrated" dotted line, meaning it was consistently too optimistic.

To understand what the AI was actually looking at to make these confident (but wrong) predictions, we analyzed the Feature Importance (Fig. 6). The results showed that the Home Odds (0.33 importance) and Away Odds (0.24 importance) were the primary drivers. These market signals vastly outweighed team performance metrics like Away Avg Shots (0.13). The AI was effectively trying to read the bookmaker's mind rather than the game itself.

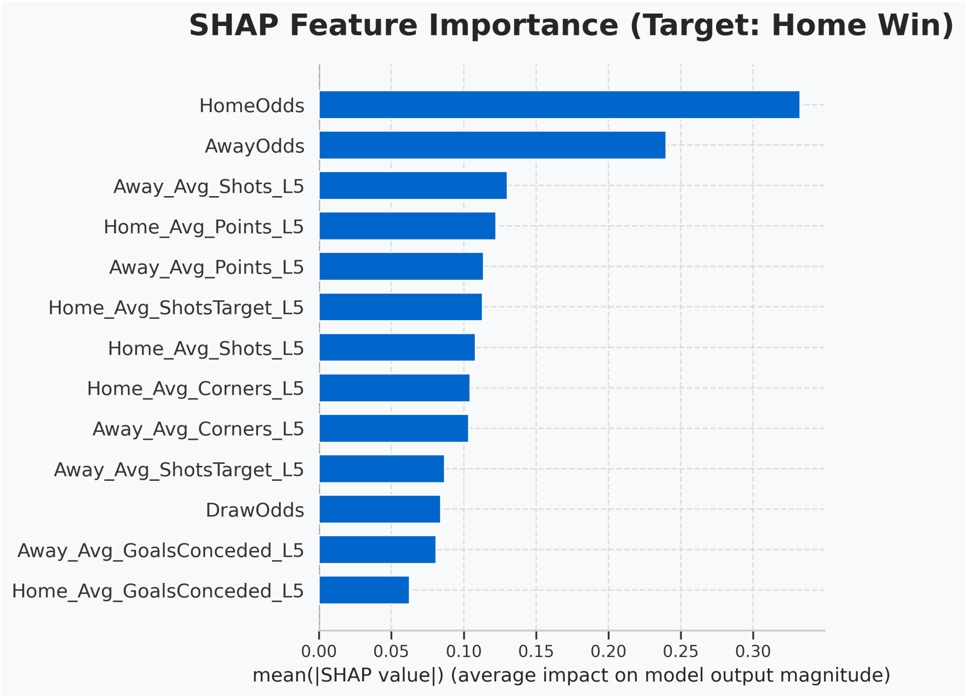


Figure 6: SHAP Feature Importance. The blue bars represent how much each data point contributed to the AI's decision. The most important feature was "HomeOdds"—the price set by the bookmaker. This suggests the AI spent more effort analyzing the market price than the actual team statistics (like Shots or Points).

4.5 Risk Assessment via Monte Carlo Simulation

Averages can be misleading. Just because the "average" simulation made a small profit doesn't mean a real person would. To see the true risk, we ran 20,000 simulations of the best performing strategy (XGBoost) to see the range of possible outcomes for a bettor.

The results (Fig. 7) were sobering. While the strategy was profitable 58.3% of the time, the risk profile reveals hidden dangers. The strategy achieved a Sortino Ratio of 0.38, but the risk profile is dangerous. As shown in Figure 7, **36.6% of all simulated lifetimes were unprofitable**. Furthermore, the 95% Value at Risk (VaR) analysis shows significant downside potential in the worst-case scenarios. Additionally, there was a 5.1% "Risk of Ruin"—meaning more than 1,000 of our simulated players went completely bankrupt.

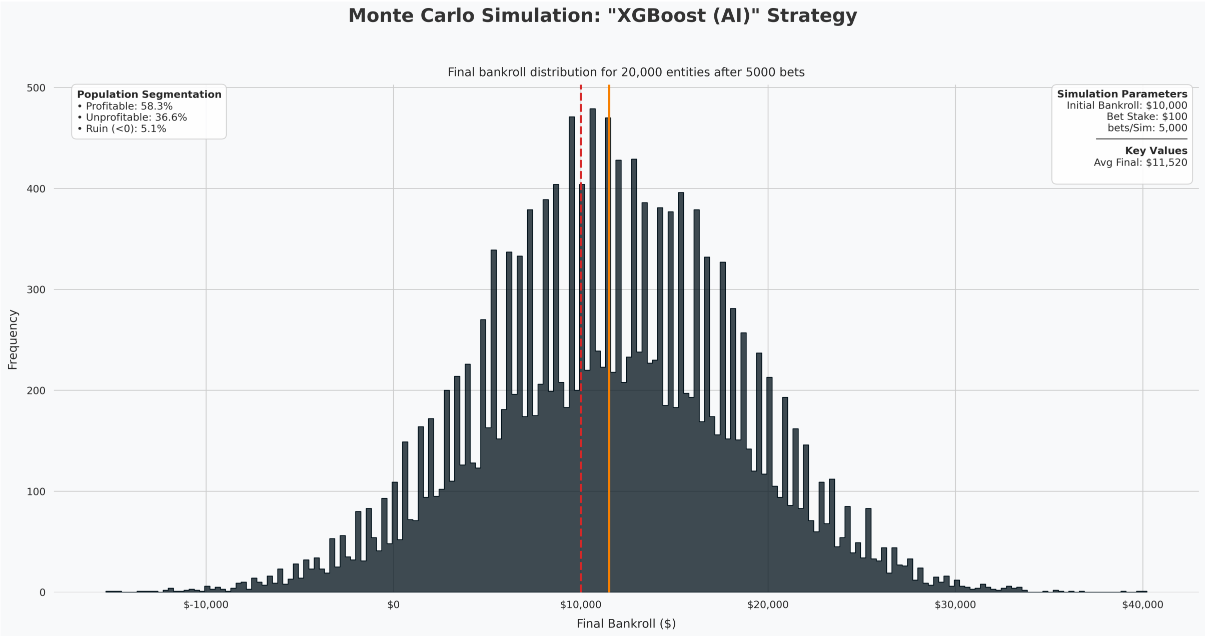


Figure 7: Monte Carlo Simulation (XGBoost). This histogram shows the final bankroll for 20,000 simulated players. While the average (orange line) is profitable at $11,520, the distribution is wide. The red dashed line shows the starting $10,000. Anyone to the left of that line lost money.

When we compare this to standard human strategies (Fig. 8), we see that the AI (green curve) shifts the outcomes slightly to the right compared to simply betting on the Favorite (blue) or the Home Team (red). It doesn't guarantee a win, but it tilts the odds slightly in your favor—just not enough to overcome the variance for everyone.

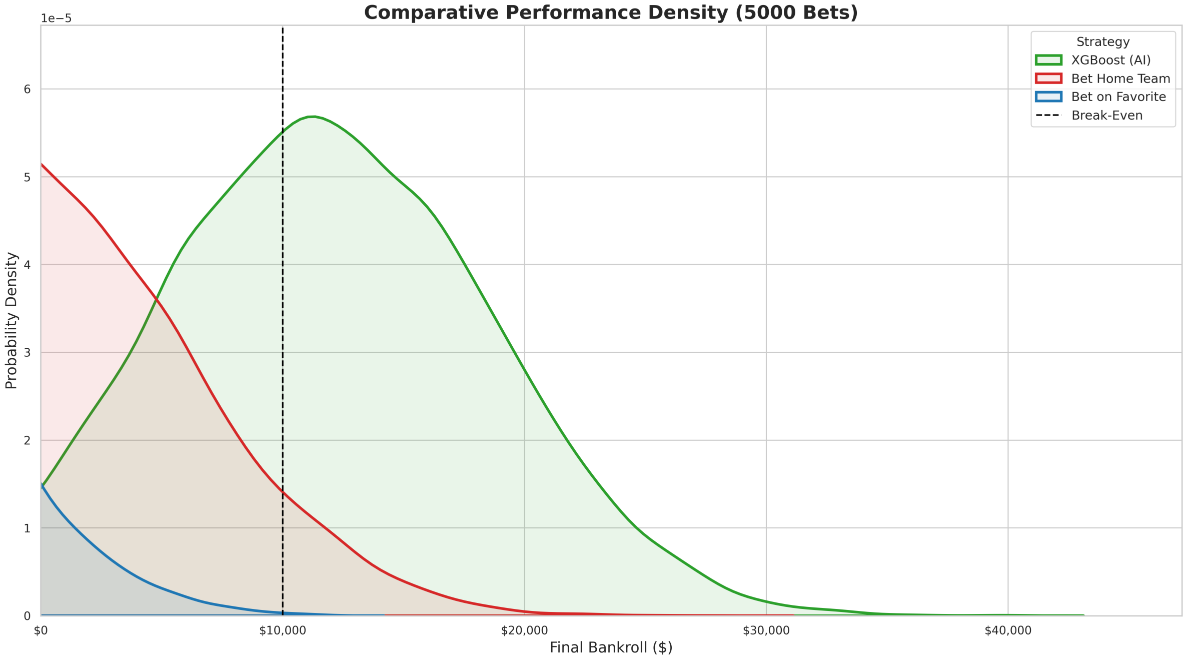


Figure 8: Comparative Performance Density. This chart overlays the AI's performance (green) against human strategies. The AI curve is taller and shifted right, meaning it produces consistent, moderate results more often than the volatile human strategies.

5 Discussion

Our findings offer a cautionary tale for the application of machine learning in competitive, human-driven markets. The failure of sophisticated models to generate a consistent profit in the modern era (post-2015) is not a failure of the algorithms themselves, but a testament to the efficiency of the market they are trying to beat.

5.1 The Evolution of the Opponent

The clear decay in performance over time supports the "Adaptive Market Hypothesis." In 2006, a standard gradient boosting algorithm could likely outpace the human compilers setting the odds. Today, those same bookmakers are almost certainly using similar, if not superior, automated systems. We are no longer playing against a human; we are playing against a mirror. The market has effectively "priced in" the very technology we are trying to use against it.

5.2 The Danger of Metric Fixation

This study also highlights a flaw in how data scientists typically evaluate models. In a classroom setting, we optimize for Accuracy or AUC (Area Under the Curve). By those standards, our models looked fine—they predicted match outcomes correctly more than half the time (approx 51%). However, in an applied setting where risk management is key, "Accuracy" is a misleading metric.

The failure of the Kelly Criterion strategy proves that **Calibration** is the more vital metric. A model that knows it is guessing (and bets small) is infinitely more useful than a model that thinks it is a genius (and bets big). The 11% calibration error we observed is the difference between a viable investment strategy and a losing one.

5.3 Strategies for Mitigation

Our results showed that the standard Kelly Criterion failed due to parameter uncertainty—the model was too confident. This aligns with findings by Baker and McHale (2013), who suggest that exact Kelly betting is perilous when probability estimates are imperfect [17]. A "Fractional Kelly" approach (betting half or quarter of the recommended amount) is often proposed as a solution to this volatility.

Furthermore, while our fundamental models struggled, other approaches exist. Kaunitz et al. (2017) demonstrated that strategies focusing purely on "consensus odds"—identifying outliers where one bookmaker is slow to update compared to the rest of the market—can still yield profits [16]. This suggests that while predicting the *game* is hard (as our study shows), predicting the *market errors* might remain viable. Conversely, Boshnakov et al. (2017) achieved success using bivariate Weibull count models, suggesting that perhaps our feature set was too simplistic and that complex parametric models still hold value [15].

6 Limitations

While this study provides robust empirical evidence regarding market efficiency, certain constraints in our scope and methodology must be acknowledged to contextualize the findings.

6.1 Data Availability and Feature Scope

Our sports betting analysis relied on a public dataset that, while historically extensive (2000–2021), was limited in feature depth. The dataset contained only match-level statistics and odds. We did not incorporate granular data points such as player-level injuries, lineup changes, or advanced metrics like Expected Goals (xG) simply because reliable historical data for these features was unavailable for the full twenty-year period. It remains an open question whether a model trained on such deep, granular data could uncover inefficiencies that our match-level models missed.

6.2 Closing Line Rigidity

Our Walk-Forward validation tested the models against the "Closing Odds"—the final price available before the match starts. The "Closing Line" is statistically proven to be the most accurate version of the market price. In a real-world scenario, professional bettors often target "Opening Odds" (days before the match), attempting to identify mispricing before the market corrects it. By testing against the Closing Line, we effectively forced our AI to beat the market at its strongest point. A model that fails against the Closing Line might still have been profitable against the softer Opening Line, though testing this hypothesis requires data that is rarely publicly available.

7 Conclusion

This study began with a simple question: Can standard machine learning algorithms outsmart a highly efficient prediction market? The answer, as it turns out, is a qualified "no"—but the reasons for that failure reveal a great deal about the state of modern data science.

By benchmarking our models against both a control group of pure noise and twenty years of real-world betting data, we established two key facts. First, our algorithms are working correctly; they successfully ignored the random noise of the control group and identified unique, statistically distinct patterns in the football data. Second, despite this technical success, they failed to generate a sustainable profit in the modern era.

The evidence points to a market that has fundamentally changed. The clear decay in performance—from consistent profits in the mid-2000s to consistent losses after 2015—suggests that the window for simple algorithmic arbitrage has closed. The "edge" that data scientists look for has been smoothed out by the efficiency of the bookmakers.

Most importantly, our results highlight a critical blind spot in how predictive models are built and tested. The fact that our models could predict the winner with respectable accuracy, yet lose money when betting with the Kelly Criterion, proves that **accuracy is not enough**. The models suffered from a significant calibration error, consistently overestimating their own certainty. In a low-margin, high-risk environment, this lack of self-awareness is fatal.

Ultimately, this research suggests that in mature, adversarial environments, the challenge is no longer about building more complex models to find better answers. The challenge is building more honest models that know when they are guessing. As markets become more efficient, the most valuable trait in an algorithm is not just its ability to predict the future, but its ability to accurately measure its own doubt.

A Appendix: The Control Experiment

As discussed in the Methodology section, we deemed it necessary to validate our models against a dataset of verified randomness before applying them to the sports betting market. This "control group" was essential to ensure that any signal found in the football data was genuine, rather than a result of the models "overfitting"—memorizing noise and mistaking it for a pattern.

A.1 Statistical Verification

First, we formally tested the data for randomness using the Augmented Dickey-Fuller (ADF) test. This is a standard statistical method used to check if a timeline of data follows a trend or if it is merely random fluctuation (stationary). The test returned a p-value of effectively zero (1.06×10^{-29}). This confirms, with high statistical confidence, that the data has no predictable structure or trend. It is pure noise.

A.2 Model Validation Test

We then trained our machine learning models on this data. To make the test rigorous, we compared two scenarios:

1. **Real Sequence:** The models tried to predict the next crash point based on the actual history.
2. **Shuffled Sequence:** We scrambled the order of the data and asked the models to predict that.

The error rates were statistically identical (Mean Absolute Error difference ¡ 0.01). This negative result is a positive finding for our study. It proves that when faced with pure noise, our algorithms correctly identify it as unpredictable.

A.3 Financial Baseline: The Mechanics of Ruin

Beyond simply testing for predictability, we wanted to establish a financial baseline for what ”randomness” looks like to a bettor. Using the house-edge statistics derived from the data (approximately 3%–7% depending on the cash-out target), we ran a Monte Carlo simulation of 2,000 players over 2,500 rounds to test standard betting heuristics.

This serves as a critical point of comparison for the main study. If our AI models in the Premier League fail, we need to know if they fail like a ”Fixed” bettor (a slow, mathematical bleed) or a ”Martingale” bettor (catastrophic variance).

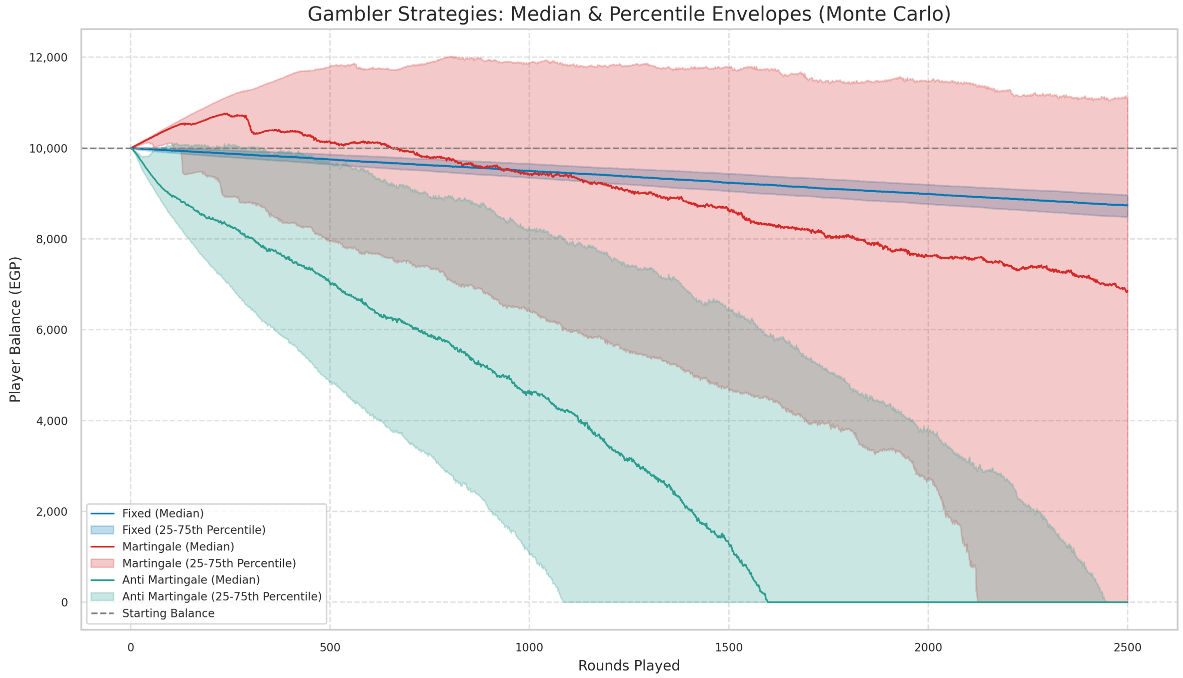


Figure 9: Monte Carlo simulation of human betting strategies on the control dataset. The ’Fixed’ strategy (Blue) visualizes the house edge as a linear decay. The ’Martingale’ strategy (Red) maintains a stable median for a time—creating an illusion of safety—but introduces massive tail risk, resulting in a 29.3% probability of total ruin within the sample period.

The simulation results (Figure 9 and Table 3) illustrate that in a purely random environment with a house edge, volatility management strategies do not alter the Expected Value. The "Martingale" system simply trades the frequency of losses for the severity of losses.

Table 3: Simulation Metrics (2,000 Players, $10,000 Start)

Strategy	**Median Final Balance**	**Probability of Ruin**
Fixed Bet	\$8,737	0.0%
Martingale	\$6,855	29.3%
Anti-Martingale	\$0	76.6%

This confirms that without a predictive edge (which the ADF test proved does not exist here), no staking method can overcome the mathematical advantage of the house.

Works Cited

[1] Grand View Research. Online gambling market size, share & trends analysis report. 2024. `https://www.grandviewresearch.com/industry-analysis/online-gambling-market`.

[2] E. F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.

[3] R. M. Griffith. Odds adjustments by American horse-race bettors. *American Journal of Psychology*, 62(2):290–294, 1949.

[4] P. W. S. Newall and D. Cortis. Are sports bettors biased toward longshots, favorites, or both? A literature review. *Risks*, 9(1):22, 2021.

[5] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[6] J. L. Kelly. A New Interpretation of Information Rate. *The Bell System Technical Journal*, 35(4):917–926, 1956.

[7] F. X. Diebold and R. S. Mariano. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

[8] Louis Chen. Football Results and Betting Odds Data of EPL. Kaggle dataset, 2021. `https://www.kaggle.com/datasets/louischen7/football-results-and-betting-odds-data-of-epl`.

[9] G. Angelini and L. De Angelis. Efficiency of online football betting markets. *International Journal of Forecasting*, 35(2):712–721, 2019.

[10] R. D. Sauer. The economics of wagering markets. *Journal of Economic Literature*, 36(4):2021–2064, 1998.

[11] S. D. Levitt. Why are gambling markets organized so differently from financial markets? *The Economic Journal*, 114(495):223–246, 2004.

[12] M. J. Dixon and S. G. Coles. Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, 46(2):265–280, 1997.

[13] E. Wheatcroft. Evaluating the impact of calibration on the profitability of betting strategies. *International Journal of Forecasting*, 36(3):1079–1090, 2020.

[14] A. Walsh and A. Joshi. Calibration is all you need: Maximizing utility in sports betting. *arXiv preprint arXiv:2401.00000*, 2024.

[15] G. N. Boshnakov, T. Kharrat, and I. G. McHale. A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458–466, 2017.

[16] L. Kaunitz, S. Zhong, and J. Kreiner. Beating the bookies with their own numbers - and how the online sports betting market is rigged. *arXiv preprint arXiv:1710.02824*, 2017.

[17] R. D. Baker and I. G. McHale. A time-varying hazard model for the exact time of goals in football. *International Journal of Forecasting*, 29(4):684–692, 2013.