Multidisciplinary Output Research For Actual and International Issue (MORFAI Journal) ISSN (e): 2808-6635
Volume 6 No. 2 (2026)

# IMPROVEMENT THE XGBOOST MODEL IN DETERMINING NBA GAME WIN DETERMINANTS: A BAYESIAN HYPERPARAMETER OPTIMIZATION AND SHAP APPROACH

**Fija Ramadhan[1]*, Ahmad Zainul Fanani[2]**
Universitas Dian Nuswantoro
E-mail: fijaramadhan10@gmail.com[1]*, a.zainul.fanani@dsn.dinus.ac.id[2]

Received : 01 December 2025
Revised : 15 December 2025
Accepted : 10 January 2026
Published : 31 January 2026
Link Publish : https://radjapublika.com/index.php/MORFAI/article/view/5028

**Abstract**

In the era of modern sports analytics, post-game team performance evaluation is often conducted subjectively. This study aims to develop an objective diagnostic model to identify the key technical–statistical factors that determine victories in the NBA, grounded in the principle of game efficiency. Using box-score statistical data from the 2004–2024 seasons, this research employs the Extreme Gradient Boosting (XGBoost) algorithm [2], optimized through the Tree-structured Parzen Estimator (TPE) method within the Optuna framework [3], to classify game outcomes. The experimental results demonstrate highly precise model performance, achieving an accuracy of 95.2% and an F1-score of 0.96. Interpretability analysis using the SHAP (Shapley Additive exPlanations) method [4] reveals that dominance in Shooting Efficiency (EFG%) and Player Impact Estimate (PIE) constitutes the absolute determinants of victory, followed by the minimization of turnovers. Furthermore, counterfactual simulations provide diagnostic insights indicating that an increase in a single statistic (e.g., +5 assists) without a corresponding improvement in shooting efficiency actually reduces the probability of winning (from 0.61 to 0.26). This finding suggests the phenomenon of "empty assists," where ball movement does not translate into effective scoring opportunities. This study contributes a performance-auditing framework that enables coaches and analysts to retrospectively evaluate the effectiveness of game strategies in an objective and data-driven manner.

***Keywords: Diagnostic Analysis, NBA, XGBoost, SHAP, Bayesian Optimization***

## INTRODUCTION

The evolution of analytics in basketball, particularly in the National Basketball Association (NBA), has transformed the paradigm of performance evaluation from intuition-based judgments to quantitative, data-driven assessments. Dean Oliver, in his seminal work, laid the foundation of modern basketball analytics by introducing the "Four Factors" as core metrics for measuring team efficiency [1]. However, with the rapid advancement of data recording technologies, Zuccolotto and Manisera emphasized that the primary challenge today is no longer data collection, but rather the extraction of hidden patterns and knowledge discovery from the massive complexity of statistical data [9]. To address this challenge, Machine Learning (ML) techniques have been widely adopted. Bunker and Thabtah noted that ML frameworks offer superior computational capabilities compared to conventional statistical approaches [6]. Nevertheless, the role of ML is not limited to outcome prediction alone. Horvat and Job highlighted that ML has evolved into a vital tool for strategic decision making in team management [10]. This perspective is reinforced by the recent study of Zhang and Gomez (2024), which demonstrated that advanced computational models can provide prescriptive insights for objectively evaluating team performance quality, going beyond mere win–loss forecasting [12].

From a technical standpoint, the Scikit-learn library is commonly employed as an industry standard for model development [7]. Among various algorithms, the Gradient Boosting technique introduced by Friedman [5] serves as a fundamental basis, which was later refined into Extreme Gradient Boosting (XGBoost) by Chen and Guestrin [2]. XGBoost is selected in this study due to its high scalability and superior performance in handling structured tabular data such as NBA box-score statistics. Despite its strength, XGBoost performance is highly sensitive to hyperparameter configurations. Manual tuning is often inefficient and suboptimal. Bergstra et al. proposed the Tree-structured Parzen Estimator (TPE) as a probabilistic optimization algorithm to overcome this limitation [11]. The implementation of this method through the Optuna framework by Akiba et al. enables automated hyperparameter search, ensuring that the model operates at its optimal performance level [3]. However, high

predictive accuracy often comes at the expense of interpretability, giving rise to the so-called black-box problem. Lundberg and Lee addressed this issue through the SHAP (Shapley Additive exPlanations) method, which provides a unified and transparent approach to explaining model outputs [4]. Such transparency is crucial for the acceptance of ML models as practical decision-support tools for coaches. A recent study by Ouyang et al. (2024) attempted to integrate XGBoost and SHAP for NBA prediction tasks [8]. Nevertheless, their work was limited by a relatively short temporal data range. This study addresses this gap by extending the analysis to a 20-season longitudinal dataset and adopting an ex-post diagnostic approach. The novelty of this research lies in the identification of the "Empty Assists" phenomenon through counterfactual analysis, which offers new strategic insights for coaches: increasing assist volume without corresponding shooting efficiency may paradoxically reduce the probability of winning. This finding underscores that not all assists contribute equally to effective offensive performance and highlights the importance of efficiency-oriented playmaking.

## LITERATURE REVIEW

### 2.1 Sports Statistics and the Four Factors

In basketball analytics, the determinants of victory are not evaluated solely from the final score, but rather from the efficiency of the underlying processes. Dean Oliver [1] formulated the concept of the Four Factors as the primary indicators of team performance. The factor with the greatest weight (40%) is Shooting, which is measured using the Effective Field Goal Percentage (eFG%). This metric is more accurate than the conventional field goal percentage because it accounts for the additional point value of three-point shots, as expressed in Equation (1):

$$\text{eFG\%} = \frac{\text{FGM} + 0.5\,.\,\text{3PM}}{\text{FGA}}$$

where FGM denotes Field Goals Made, 3PM denotes Three-Point Made, and FGA denotes Field Goals Attempted.

Beyond offensive efficiency, other crucial factors include minimizing Turnovers (loss of ball possession), dominance in Rebounds, and the frequency of Free Throws.

### 2.2 Extreme Gradient Boosting (XGBoost)

XGBoost is an ensemble algorithm based on decision trees that follows the principle of gradient boosting. It builds models in an additive manner, where each new tree is trained to correct the residual errors of the previous trees [2].

The main advantage of XGBoost lies in its objective function, which combines a loss function with a regularization term ($\Omega$) to prevent overfitting. Mathematically, the objective function to be minimized is given in Equation (2):

$$\mathcal{L}(\phi) = \sum_{i} l\,(\hat{y}_i, y_i) + \sum_{k} \Omega\,(f_k)$$

where:

- $l(\hat{y}_i, y_i)$ is a differentiable loss function (commonly Log-Loss for binary classification) that measures the discrepancy between the model prediction ($\hat{y}_i$) and the actual label ($y_i$)
- $\Omega(f_k)$ is the regularization penalty function that controls the complexity of each tree ($f_k$), ensuring the model maintains good generalization performance on unseen data.

### 2.3 Bayesian-Based Hyperparameter Optimization (Optuna)

The performance of XGBoost is highly sensitive to hyperparameter settings (such as learning rate, tree depth, and subsampling ratio). Conventional search methods such as Grid Search are often computationally inefficient because they exhaustively explore all combinations in the parameter space. This study employs a Bayesian Optimization approach using the Optuna framework. Optuna implements the Tree-structured Parzen Estimator (TPE) algorithm to model the probability $P(x|y)$ based on the results of previous evaluations [3]. Through this probabilistic approach, the algorithm can intelligently select the most promising hyperparameter configurations for subsequent iterations. Consequently, convergence toward the optimal solution can be achieved more efficiently compared to random or exhaustive search methods.

### 2.4 Model Interpretability with SHAP

One of the major challenges in using complex ensemble models such as XGBoost is their black-box nature. To provide diagnostic transparency, this study employs the SHAP (Shapley Additive exPlanations) method.

SHAP is grounded in the concept of Shapley values from Cooperative Game Theory. It computes the marginal contribution of each feature to the model's final prediction. Mathematically, the attribution value of a feature $\phi_i$ is calculated as shown in Equation (3):

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \backslash i)\right]$$

where:

- $M$ is the total number of features,
- $z'$ is a subset of features,
- the term $[f_x(z') - f_x(z' \backslash i)]$ represents the marginal impact of adding feature $i$ to the model prediction [4].

The additive property of SHAP ensures consistency in interpretation, enabling coaches and analysts to precisely understand which statistical variables act as the primary determinants of wins or losses.

## METHOD

### 3.1 Research Framework

This study adopts a structured Machine Learning workflow to ensure the reproducibility and validity of the results. The research stages include data acquisition, feature engineering, automated model optimization, and diagnostic interpretation. The overall workflow is visually illustrated in Figure 1.

### 3.2 Dataset Collection

The dataset used in this study was obtained from the official nba_api repository, which covers all NBA regular-season games from 2004 to 2024. The raw dataset consists of standard box-score statistics (Points, Rebounds, Assists, etc.). The target label $(y)$ is binary, where a value of 1 represents a home team victory (Home Win) and 0 represents an away team victory (Away Win).

### 3.3 Preprocessing and Feature Engineering

This stage aims to transform raw data into an informative format suitable for the model.

- **Data Cleaning:** Rows containing missing values or duplicate entries were removed to ensure data quality and consistency.
- **Delta Feature Transformation (Δ):** Since a basketball game is a zero-sum competition, the absolute statistics of a single team are insufficient to describe competitive advantage. Therefore, a differential feature transformation was applied for each statistical variable j in game i:

$$X_{i,j}^{(\Delta)} = X_{i,j}^{(\text{Home})} - X_{i,j}^{(\text{Away})}$$

A positive value indicates a statistical advantage for the home team.

- **Data Splitting:**The dataset was divided into a training set (80%) and a testing set (20%). The split was performed chronologically rather than randomly to prevent data leakage and to ensure that the model is evaluated on games that occurred after the training period.

### 3.4 Model Architecture and Optimization

The classification algorithm employed in this study is XGBoost (Extreme Gradient Boosting). To avoid subjectivity in parameter selection, Bayesian Optimization was applied using the Optuna library.
The hyperparameter search space was defined as follows:

- learning_rate: [0.01, 0.3] (model learning rate)
- max_depth: [3, 10] (maximum depth of decision trees)
- n_estimators: [100, 500] (number of trees)
- subsample: [0.5, 1.0] (proportion of samples used for each tree)
- colsample_bytree: [0.5, 1.0] (proportion of features used per tree)

The Tree-structured Parzen Estimator (TPE) algorithm implemented in Optuna performs 20 optimization trials to minimize the Log-Loss function and identify the most optimal hyperparameter configuration.

### 3.5 Evaluation Scenario and Performance Interpretation

Model performance was evaluated using standard classification metrics: Accuracy, Precision, Recall, and F1-Score. Further validation was conducted using the Confusion Matrix to examine the distribution of prediction errors (Type I and Type II errors).
After obtaining the best-performing model, interpretability analysis was carried out using two approaches:

- **Global Analysis (SHAP Summary Plot):** Identifies which features have the greatest average impact on overall win predictions.
- **Local Analysis (Counterfactual Simulation):** Conducts "what-if" experiments on specific game samples. Selected input variables (e.g., Assists) are manipulated to measure the elasticity of winning probability in response to changes in game strategy.

## RESULTS AND DISCUSSION

### 4.1 Data Description and Class Separation (PCA)

The initial stage of this study involves transforming NBA game statistics into differential (Delta) values to represent the relative advantage between competing teams. Visualization using Principal Component Analysis (PCA) is then performed to examine the spatial distribution of the data and to observe the degree of class separation between winning and losing outcomes.

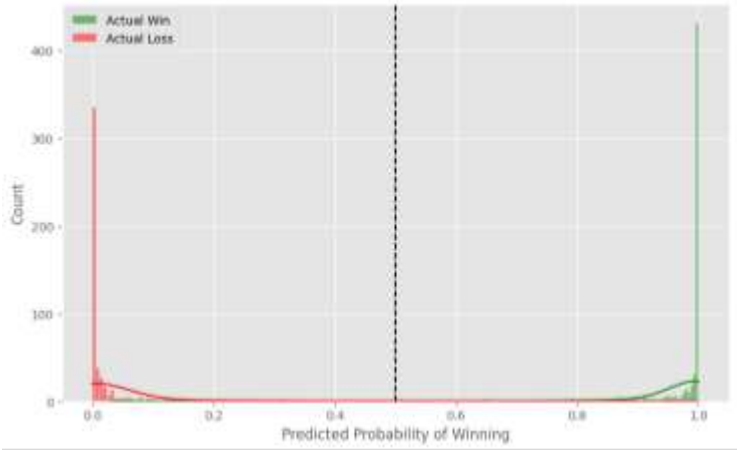
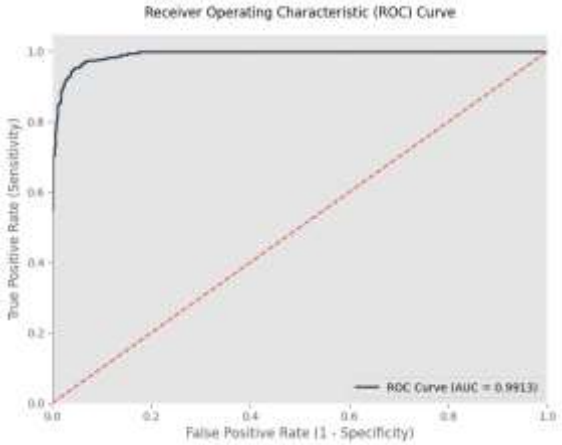



Figure 1. Visual Segregation between Win and Loss Classes Using PCA.

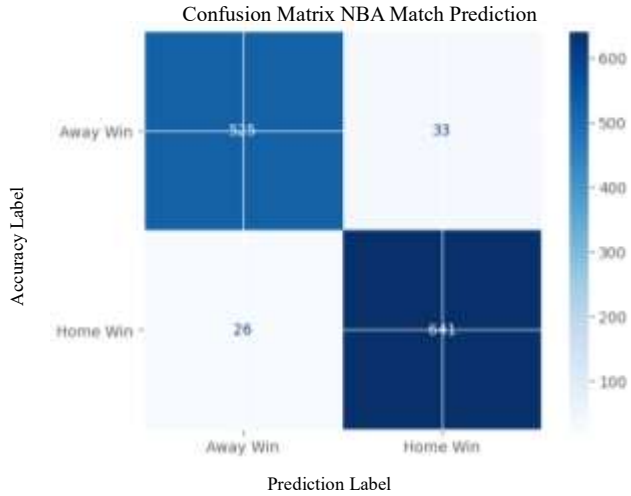
The PCA plot in Figure 4.1 shows a relatively clear separation between the winning class (1) and the losing class (0). Although some overlap exists in the central region, the overall distribution pattern confirms that game statistics possess an underlying structure that can be mathematically captured by non-linear models.

### 4.2 Performance Evaluation of the XGBoost Model with Optuna Optimization

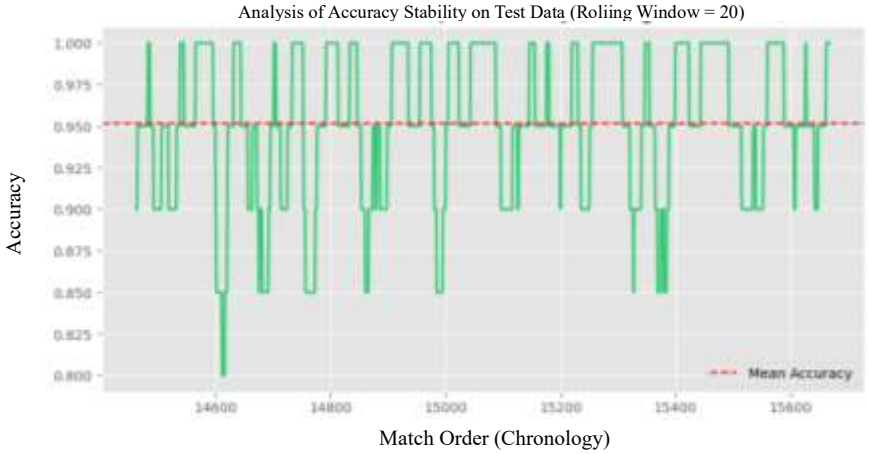
The XGBoost model optimized using Bayesian Optimization through the Optuna framework demonstrates very high and stable performance.




Validation of the Model's Discriminative Ability with an AUC of 0.9913. An AUC score of 0.9913 (Figure 4.2) indicates that the model has an almost perfect capability to discriminate between game outcomes. This result is further supported by the Confusion Matrix presented below, which shows a very low level of misclassification, confirming the robustness and reliability of the model's predictive performance.



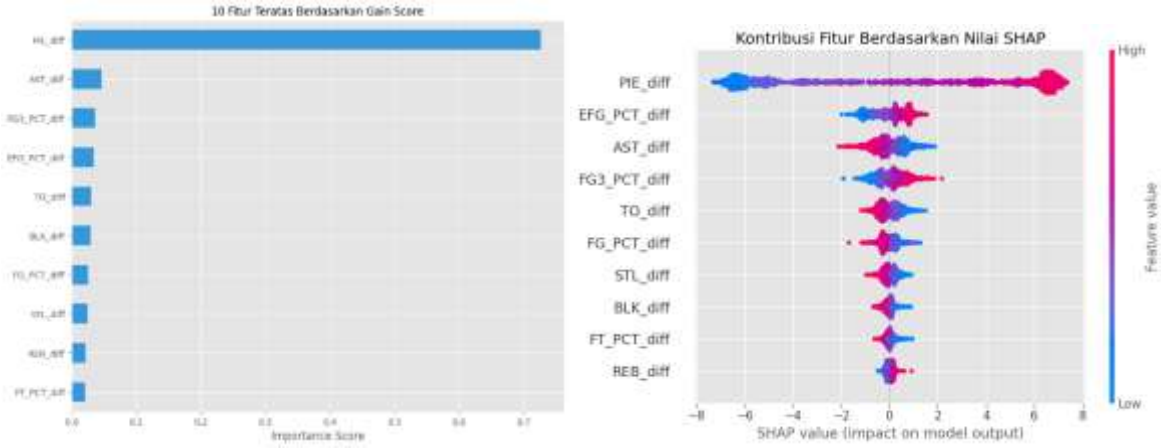

Classification Precision of the Model on the Test Data (Accuracy of 95.2%). Based on Figure 4.3, the model makes only 59 errors out of a total of 1,225 test samples, resulting in an overall accuracy of 95.2%. Further validation is conducted using a rolling window analysis to ensure the model's consistency across different seasons.




Consistency of Model Accuracy Across Seasons Using a Rolling Window. As shown in Figure 4.4, the model's accuracy remains stable with an average around 95%, demonstrating that the model does not experience significant performance degradation over time in a chronological setting.

### 4.3 Analysis of Win Determinants

The core of the model's transparency lies in identifying the variables that most strongly influence winning outcomes using Gain Score and SHAP analysis.

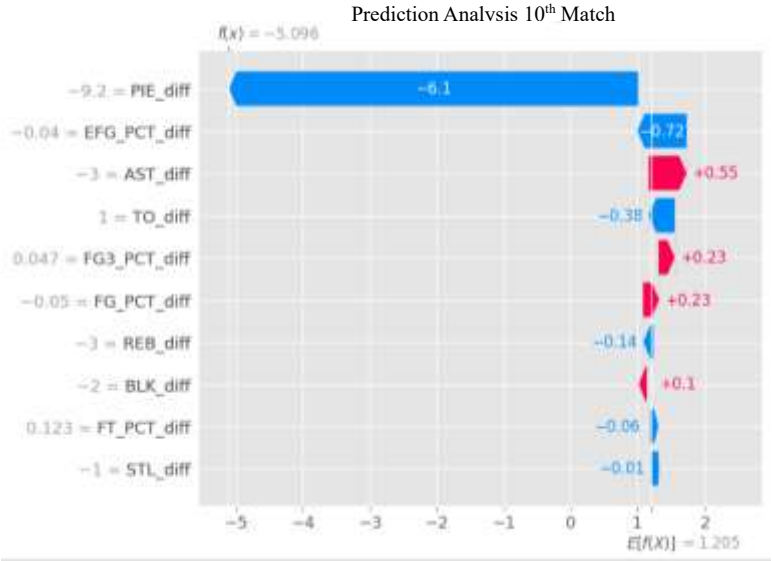



Based on Figures 4.5 and 4.6, the feature PIE_diff (Player Impact Estimate) emerges as the most dominant determinant of victory. It is followed by shooting efficiency (EFG_PCT_diff) and the number of assists (AST_diff). In contrast, the variable TO_diff (Turnovers) exhibits a strong negative impact (indicated by the red points on the left side of the zero axis in the SHAP plot), implying that the higher the number of self-inflicted errors, the lower the probability of winning.

### 4.4 Local Interpretation and Strategic Simulation

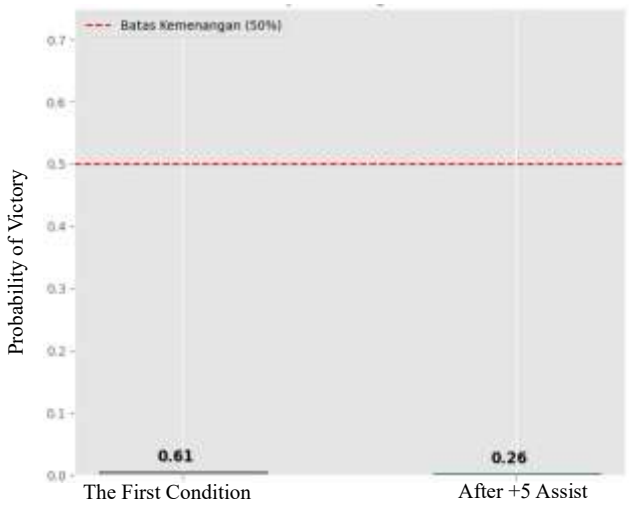
This study is able to diagnose individual game outcomes through the decomposition of contributing factors.

Prediction Analysis 10th Match




The figure above visualizes the decomposition of the model's decision for a specific game (Index 10). In this plot, the X-axis represents the log-odds value, where the model starts from the baseline expectation E[f(x)] and ends at the final prediction score f(x)= -5.096. This large negative value indicates that the model is highly confident that the team will lose the game.

Simulation : Impact of Assist Improvement (+5)




The simulation results in Figure 4.8 reveal an intriguing anomaly in which adding +5 Assists under the given game conditions actually decreases the probability of winning (from 0.61 to 0.26). This finding diagnoses the phenomenon of "empty assists," where an increase in passing volume without being accompanied by efficient scoring execution (low PIE) is ineffective in altering the game outcome.

## CONCLUSION

This study successfully achieves its primary objective of developing an objective diagnostic model to identify the determinants of NBA victories. Based on empirical evaluation, the application of the Extreme Gradient Boosting (XGBoost) algorithm [2], optimized using the Bayesian TPE approach [3], is proven to be highly effective in capturing complex winning patterns, delivering precise performance with an accuracy of 95.2% and an AUC value of 0.9913. The stability of the model is validated across competitive seasons, confirming that the statistical variables employed exhibit strong predictive consistency in line with the theoretical foundation of the Four Factors [1]. From a diagnostic perspective, this study concludes that winning is not determined merely by the accumulation of volume-based statistics, but rather by game efficiency. Feature importance analysis using SHAP [4] quantitatively demonstrates that Player Impact Estimate (PIE) and Effective Field Goal Percentage (eFG%) are the dominant indicators of victory, while Turnovers exert a significant negative impact. A crucial finding from the counterfactual simulation reveals the phenomenon of "empty assists," in which an increase in the number of assists (+5) without accompanying efficiency in scoring execution drastically reduces the probability of winning (from 0.61% to 0.26%). As a future service development plan, this framework is recommended to be implemented as an ex-post performance auditing tool for coaches. For further research, it is suggested that the model should not be limited to box-score data alone, but be integrated with spatial data (player tracking) to capture off-ball movement dynamics. Additionally, expanding the scope of analysis to other leagues is encouraged in order to test the generalizability of these win determinants across different competitive contexts.

## REFERENCES

[1] D. Oliver, Basketball on Paper: Rules and Tools for Performance Analysis. Washington, D.C.: Potomac Books, Inc., 2004.
[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.
[3] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 2019, pp. 2623–2631.
[4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 2017, pp. 4765–4774.
[5] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[6] R. P. Bunker and F. Thabtah, "A Machine Learning Framework for Sport Result Prediction," Applied Computing and Informatics, vol. 15, no. 1, pp. 27–33, 2019. doi: 10.1016/j.aci.2017.09.005.
[7] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[8] Y. Ouyang et al., "Integration of machine learning XGBoost and SHAP models for NBA game outcome prediction and quantitative analysis methodology," PLoS ONE, vol. 19, no. 7, p. e0307478, Jul. 2024.
[9] P. Zuccolotto and M. Manisera, Basketball Data Science: With Applications in R. CRC Press, 2020.
[10] T. Horvat and J. Job, "The use of machine learning in sport outcome prediction: A review," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 5, p. e1380, 2020.
[11] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for Hyper-Parameter Optimization," in Advances in Neural Information Processing Systems (NIPS), 2011, pp. 2546–2554.
[12] L. Zhang and M. A. Gomez, "Machine Learning for Basketball Game Outcomes: NBA and WNBA Leagues," Algorithms, vol. 13, no. 10, p. 230, Oct. 2024.