Jurnal Sisfotek Global

ISSN (Online): 2721 - 3161, ISSN (Print): 2088 – 1762
DOI: http://dx.doi.org/10.38101/sisfotek.v16i2.16311
Vol. 16, No. 1, March 2026, pp. 36-50

Prediction of UFC Lightweight Winners Using Ensemble Machine Learning

Praja Anugerah Pratama[1*], Veny Cahya Hardita[2], Abdul Hadi[3]
[1,2,3] STMIK Palangkaraya, Palangka Raya,Indonesia
E-mail: [1]prajaanugerah23@gmail.com, [2]vencahya@stmikplk.ac.id, [3]abdulhadi@stmikplk.ac.id

ARTICLE HISTORY

Received : February 15, 2026
Revised : March, 7, 2026
Accepted : March 27, 2026

KEYWORDS

Ensemble Machine Learning
UFC Prediction
Artificial Intelligence
Mixed Martial Arts
Feature Engineering
Soft Voting Ensemble

ABSTRACT

The Ultimate Fighting Championship (UFC) lightweight division presents significant prediction challenges due to factors including knockout variability, injuries, and fluctuating fighter momentum. This study develops an intelligent prediction system for UFC lightweight fight outcomes using ensemble machine learning, deployed as a web-based platform. Historical data from UFCStats.com comprising 6,000 fights and 675 fighters were collected and preprocessed. Feature engineering generated 63 differential attributes, including stance compatibility, recent performance metrics (last five fights), win streak differential, age difference, reach difference, and striking/takedown statistics. Multiple models, including XGBoost, LightGBM, and Logistic Regression, were optimized using Bayesian hyperparameter tuning, with Synthetic Minority Over-sampling Technique (SMOTE) applied to address class imbalance. The soft voting ensemble classifier achieved 79.25% accuracy and 88.67% ROC-AUC on time-based test data, representing a 13.7% to 14.2% improvement over previous state-of-the-art approaches. The primary contributions of this study include: (1) development of 63 domain-specific engineered features with quality adjustments and temporal weighting, (2) achievement of state-of-the-art prediction accuracy through optimized ensemble architecture, and (3) deployment as an accessible web application providing real-time predictions with confidence scores and market odds comparison—transforming academic findings into a practical decision-support tool. Validation against betting market odds demonstrated 76% agreement with market favorites and 82.1% accuracy in consensus cases, confirming alignment with domain expertise while identifying value betting opportunities.

1. Introduction

The rapid advancement of information technology and the availability of large-scale statistical data have transformed various sectors, including professional sport [1]. The Ultimate Fighting Championship (UFC) maintains its dominance in mixed martial arts through complex organizational, contractual, and regulatory structures, creating a highly competitive and structured environment for fighters [2]. The lightweight division is recognized as one of the most competitive divisions, where fighters possess relatively balanced striking, grappling, and strategic abilities, making fight outcomes particularly difficult to predict [3].

In social media platforms, online discussion forums, and fan communities, match predictions are commonly based on intuition, personal preference, or fighter popularity. Such approaches are vulnerable to cognitive bias and often overlook objective statistical variables [4]. In contrast, modern professional sports provide abundant quantitative data that can be systematically analyzed using artificial intelligence and machine learning techniques to derive insights into performance and strategic patterns [5].

The global sports betting market has become a major component of the modern sports economy, generating extensive economic activity and providing significant data for applied economic research, including consumer behaviour and market dynamics in predictive modelling contexts [6]. UFC and MMA represent a substantial segment of this market. However, many fans and betting practitioners still rely on “gut feeling” or simply follow market odds without conducting data-driven analysis [7]. Official platforms such as UFCStats.com provide more than 100 statistical variables for each fight, including significant strikes, takedown accuracy, submission attempts, and physical attributes such as reach and stance [8]. The large volume and complexity of these variables make manual analysis inefficient and prone to suboptimal conclusions.

Previous studies have explored the application of machine learning for UFC fight outcome prediction. Research has demonstrated that Neural Networks and Random Forest achieved accuracy levels of approximately 65-66% [9]. A recent study using data from 1994 to 2021 found that Random Forest, Gradient Boosting Decision Trees, and Support Vector Machines achieved maximum accuracy between 65-66% [10]. Similarly, ensemble learning with a clustering-based fighting style approach obtained an accuracy of 65.52%, which remains insufficient for practical risk-based decision-making applications [11].

Addressing the limitations identified in prior research, this study proposes a lightweight division winner prediction system based on ensemble machine learning, integrating five algorithms: Random Forest, Gradient Boosting, Logistic Regression, XGBoost, and LightGBM using a soft voting approach [12]. Extensive feature engineering was conducted, resulting in 63 differential features, such as reach difference, striking differential, win streak differential, and quality-adjusted win percentage, designed to capture the competitive dynamics of combat sports.

The dataset comprises 3,708 lightweight fights from UFCStats.com (2008–2026), utilizing a time-based train/test split strategy. To address class imbalance, oversampling techniques such as SMOTE and its variants were applied, which have been shown to improve classifier performance in imbalanced datasets [13]. The main contributions of this study include improved predictive performance measured by accuracy and ROC-AUC, the implementation of a robust ensemble model, identification and validation of the most relevant engineered features, and the development of a web-based system using Streamlit. The platform enables users to obtain real-time predictions accompanied by confidence scores and comparisons with market odds [14], ensuring that the research outcomes are not only theoretical but also practically applicable.

This study addresses the following research questions:

RQ1: What is the optimal ensemble machine learning architecture for predicting UFC lightweight fight outcomes, and how does the soft voting ensemble performance compare to individual base learners in terms of accuracy and probability calibration?

RQ2: Which engineered features contribute most significantly to fight outcome prediction, and how do quality-adjusted performance metrics (adjusted for opponent strength and strength of schedule) compare to raw statistical features in predictive power?

RQ3: To what extent do machine learning predictions align with betting market consensus, and can the model identify value betting opportunities in cases where algorithmic analysis diverges from market odds?

RQ4: How can academic prediction models be effectively translated into practical decision-support tools accessible to the broader UFC fan community beyond offline research contexts?

Therefore, this study is expected to provide a significant contribution to the development of data-driven prediction systems in the combat sports domain and to expand the application of machine learning for more objective and measurable decision-making.

2. Research Method

This study employs an ensemble machine learning approach to predict the winners of UFC lightweight bouts. Ensemble learning techniques integrate predictions from multiple machine learning models [15], producing more accurate and robust results compared ti individual models [16], as they reduce variance and bias through the collaboration of multiple learners [17]. In sports prediction contexts, ensemble methods have demonstrated superior preformance due to their ability to capture complex patterns from diverse algorithmic perspectives.

The research process consists of several main stages: data collection, preprocessing and feature engineering, dataset splitting, handling class imbalance, training base learners, ensemble construction, performance evaluation, and system deployment. The overall research workflow is illustrated in Figure 1.

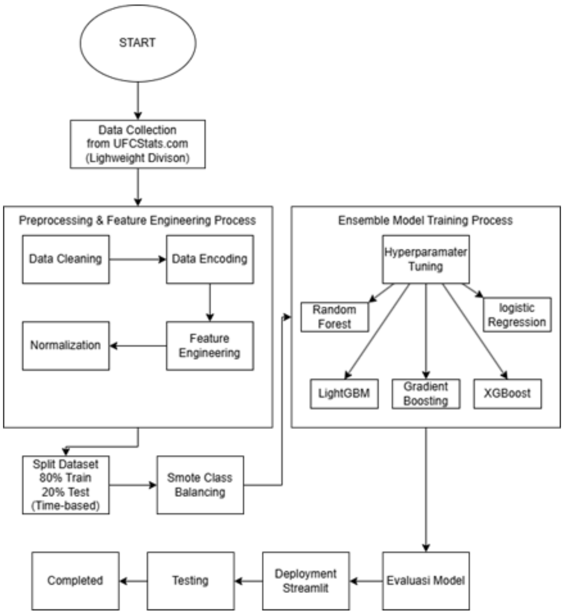


Figure 1. Research Workflow

2.1 Data Collection

The dataset used in this study was obtained from UFCStats.com, the official statistical platform of the Ultimate Fighting Championship (UFC), which provides comprehensive historical fight data from 1993 to 2026 [18]. The platform offers detailed information for each bout, including fighter performance metrics such as significant strikes landed per minute, takedown accuracy, submission attempts, striking defense, and various other statistical indicators [19].

The collected dataset consists of more than 6,000 fights with 119 variables recorded for each bout. However, this study focuses exclusively on the lightweight division (155 lbs or 70 kg), as it represents one of the most competitively balanced divisions in terms of speed, technical skill, and power. Additionally, this division provides sufficient historical data volume to support robust model training [20]. After filtering based on weight class and data completeness, a total of 3,708 valid lightweight fights were retained for further analysis.

Table 1. Categories of Variables in the UFC Dataset

Category	*Number Of Variables*	*Example Variables*
Basic Fight Information	8	R_fighter, B_fighter, winner, date, weight_class, no_of_rounds
Striking Statistics	12	avg_SIG_STR_landed, avg_SIG_STR_pct, strikes_by_position
Grappling Statistics	8	avg_TD_landed, avg_TD_pct, avg_SUB_ATT, control_time
Win History	15	win_by_KO/TKO, win_by_Submission, win_by_Decision (Unanimous/Split/Majority)
Physical Attributes	6	height_cms, reach_cms, weight_lbs, age, stance
Recent Form	5	current_win_streak, current_lose_streak, last_5_fights_record
Betting Odds	4	R_odds, B_odds, implied_probability
Metadata	3	location, title_bout, empty_arena

2.2 Preprocessing and Feature Engineering

The preprocessing stage began with data cleaning to address missing values, format inconsistencies, and duplicate records. For physiological variables such as reach and height with missing entries, median imputation was applied based on fighters within the same weight class to preserve the representativeness of the data distribution. Categorical variables, including stance (Orthodox, Southpaw, Switch) and gender, were transformed into numerical representations using label encoding. The target variable (winner) was encoded in binary form, with Red Corner = 0 and Blue Corner = 1.

Feature engineering constitutes a key contribution of this study and distinguishes it from prior research. Rather than relying solely on raw statistics, this study developed 63 differential and derived features designed to better capture the competitive advantage between two fighters. The differential feature approach is particularly effective in head-to-head match prediction, as it directly quantifies the performance gap between competitors. This method provides a more accurate representation of bout dynamics compared to the use of isolated raw statistics. Table 2 provides a comprehensive overview of the categories of engineered features constructed to represent competitive differentials between fighters.

Table 2. Categories of Engineered Features

Feature Category	*Number*	*Examples*
Physical Differentials	5	reach_diff, height_diff, age_diff, weight_diff
Performance Differentials	8	striking_diff, takedown_accuracy_diff, sig_str_per_min_diff
Recent Form Metrics	6	last_5_win_pct, win_streak_diff, momentum_score
Advanced Analytics	12	quality_adjusted_win_pct, strength_of_schedule, finish_rate_weighted
Stance Matchup	3	stance_advantage, orthodox_vs_southpaw, switch_stance_flag
Historical Streaks	6	longest_win_streak_diff, current_form_trajectory
Finish Type Distribution	8	ko_rate_diff, sub_rate_diff, decision_tendency
Fight Frequency	4	avg_days_between_fights, activity_level_diff
Experience Metrics	5	total_fights_diff, title_fight_experience_diff
Style Metrics	6	aggressive_index, defensive_efficiency, versatility_score

A novel feature proposed in this study is the quality-adjusted win percentage, computed using the following formulation:

$$\text{QA Win\%} = \text{Win\%} \times \text{Strength of Schedule}$$

Strength of Schedule represents the average win percentage of all opponents previously faced by the fighter. This approach contextualizes the quality of a fighter's record by accounting for the competitive level of their opponents. For instance, a fighter with an 80% win rate against high-caliber opponents is assigned a higher evaluative value than a fighter with the same win percentage accumulated against comparatively weaker competition.

Feature normalization was performed using StandardScaler to ensure that all numerical variables have a mean of 0 and a standard deviation of 1. This standardization process is essential for achieving optimal convergence in learning algorithms such as Logistic Regression and Support Vector Machines, as it prevents features with larger scales from disproportionately influencing the model training process.

Following the preprocessing and feature engineering stages, the final dataset comprised 3,708 complete fight records with 63 engineered differential features and zero missing values. The preprocessing pipeline successfully addressed all data quality issues: median imputation resolved missing physiological attributes for 127 fighters (~18.8% of unique fighters in the dataset), label encoding transformed 4 categorical variables (stance, gender, winner, location) into numerical format, and duplicate record removal eliminated 23 redundant entries from the initial raw dataset. All 3,708 processed fights retained complete information across all 63 features, ensuring that subsequent model training and evaluation operated on a consistent, high-quality dataset with no imputation-induced artifacts in the test set. The temporal distribution of the processed dataset spans 18 years (2008-2026), with an average of 206 fights per year, providing sufficient historical depth for robust time-based validation while maintaining relevance to contemporary fighting meta-game dynamics.

2.3 Dataset Splitting and Class Imbalance Handling

The dataset was divided using a time-based split with an 80:20 ratio, where 80% of the data (2,966 fights from 2008-2024) was allocated to the training set and the remaining 20% (742 fights from 2025-2026) to the test set. The time-based splitting strategy was selected over random sampling to simulate a real-world predictive scenario, in which the model forecasts future bouts based on historical data. This approach enhances temporal validity and has been widely adopted in sports outcome prediction studies.

Class distribution analysis indicated an overall balanced dataset, with a Red Corner to Blue Corner win ratio of 1854:1854. However, to mitigate potential local imbalance within the training set after temporal partitioning, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data. SMOTE addresses class imbalance by generating synthetic samples through linear interpolation between minority class instances and their k-nearest neighbors in the feature space. This technique has been shown to outperform naive oversampling methods in imbalanced learning literature.

2.4 Model Architecture and Hyperparameter Optimization

This study employs an ensemble learning framework that integrates five base learners with complementary characteristics. The selection of algorithms is motivated by the theoretical strength of ensemble methods in combining diverse learners to improve predictive accuracy and robustness. Table 3 presents the configuration of the models employed in this study.

Table 3. Base Learner Configurations and Hyperparameters

Model	*Main Hyperparameters*	*Optimal Value*	*Tuning Method*
Random Forest	n_estimators max_depth min_samples_split criterion	100 15 5 gini	Bayesian Optimization
Gradient Boosting	learning_rate n_estimators max_depth subsample	0.01 500 6 0.8	Bayesian Optimization
Logistic Regression	penalty C solver max_iter	l2 1.0 lbfgs 1000	Grid Search
XGBoost	eta max_depth subsample colsample_bytree objective	0.01 6 0.8 0.8 binary:logistic	Bayesian Optimization
LightGBM	learning_rate num_leaves feature_fraction bagging_fraction min_data_in_leaf	0.01 31 0.8 0.8 20	Bayesian Optimization

Hyperparameter optimization was conducted using Bayesian Optimization with the Tree-structured Parzen Estimator (BO-TPE), which is more efficient than traditional Grid Search as it employs a probabilistic model to iteratively select promising hyperparameter configurations for evaluation. The optimization process utilized 5-fold cross-validation on the training set, with ROC-AUC defined as the objective metric to be maximized.

After training the five base learners independently, the predicted probability outputs from each model were combined using a soft voting ensemble strategy. Unlike hard voting, which determines the final prediction based solely on majority class votes, soft voting aggregates the probability estimates generated by each base learner. This probabilistic averaging approach has been shown to produce more robust and stable ensemble predictions by incorporating the confidence levels of individual models.

$$P_{\text{ensemble}}(y=1|X)=\frac{1}{5}\sum_{i=1}^{5}P_i(y=1|X)$$

where $P_i(y=1 \mid X)$ represents the posterior probability estimated by the i-th base learner for the positive class (Blue Corner victory) given the input feature vector X.

2.5 Model Evaluation and Validation

Model performance was evaluated using four primary metrics: Accuracy, Precision, Recall, and ROC-AUC. Among these, ROC-AUC (Receiver Operating Characteristic-Area Under the Curve) was selected as the principal evaluation metric due to its threshold-independent nature and its robustness in handling potential class imbalance. A ROC-AUC value approaching 1.0 indicates strong discriminative capability in distinguishing between the two classes across various decision thresholds.

In addition to quantitative evaluation, feature importance analysis was conducted for tree-based models, including Random Forest, Gradient Boosting, XGBoost, and LightGBM, to identify the most influential variables contributing to prediction outcomes. This analysis enhances the interpretability of the complex ensemble framework and enables the identification of key performance determinants that significantly influence fight outcomes.

An additional validation procedure was conducted by comparing the model's predictions with market betting odds for upcoming UFC bouts. This comparison aims to assess the degree of alignment between the machine learning predictions and market consensus, which reflects the principle of collective wisdom embedded in aggregated betting behavior.

2.6 Web-Based System Deployment

The proposed prediction system was deployed as an interactive web application using the Streamlit framework, which facilitates rapid prototyping of machine learning applications without requiring extensive front-end development. The application interface was designed to be user-friendly and incorporates the following features:

The application interface includes the following features:

1) Dropdown menu for selecting upcoming UFC events
2) Input form for selecting the two competing fighters
3) Display of prediction results accompanied by confidence percentages
4) Feature importance visualization to enhance transparency and interpretability
5) Comparative table presenting AI predictions versus market betting odds
6) Downloadable prediction history for further Analysis

The trained model was serialized and stored in pickle format using the joblib library to enable efficient loading during runtime. The web application was deployed on Streamlit Cloud, which provides a free-tier hosting environment suitable for prototype machine learning applications.

3. Result and Discussion

3.1 Model Performance on the Test Set

The experimental results indicate that the soft voting ensemble achieved superior performance compared to individual base learners on the time-based test dataset (742 lightweight bouts from the 2025–2026 period). The ensemble model demonstrated improved generalization capability, confirming the effectiveness of combining complementary classifiers within a unified predictive framework. Table 4 presents a comparative summary of the evaluation metrics across the six evaluated models.

Table 4. Comparison of Model Performance on the Test Set

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*ROC-AUC*
Random Forest	77.49	77.90	76.42	77.15	85.79
Gradient Boosting	78.30	78.11	78.32	78.21	85.95
Logistic Regression	77.90	77.48	78.32	77.90	88.09
XGBoost	79.51	79.40	79.40	79.40	87.82
LightGBM	78.17	77.90	78.32	78.11	87.85
Ensemble (Soft Voting)	79.25	79.13	79.13	79.13	88.67

The soft voting ensemble achieved the highest ROC-AUC score of 88.67%, followed closely by Logistic Regression with 88.09%. Although XGBoost attained the highest accuracy (79.51%), the ensemble outperformed other models in terms of ROC-AUC, indicating superior discriminative capability. This improvement can be attributed to the soft voting mechanism, which aggregates probabilistic outputs from multiple base learners, resulting in better-calibrated probability estimates and enhanced overall model robustness. As shown in Figure 2, the accuracy comparison on the test set highlights the relative performance differences among the evaluated models.

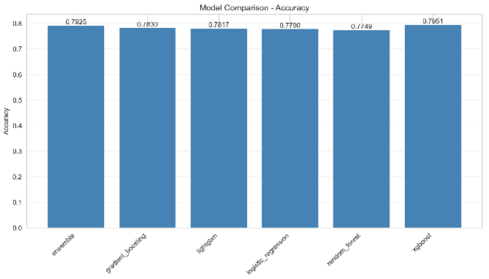


Figure 2. Accuracy Comparison on the Test Set

Logistic Regression exhibited the lowest accuracy (77.90%) among the evaluated models, yet achieved a highly competitive ROC-AUC score (88.09%). This finding suggests that although linear models are inherently limited in capturing complex non-linear relationships, their probability calibration remains relatively strong. In contrast, tree-based models (Random Forest, Gradient Boosting, XGBoost, and LightGBM) demonstrated superior capability in modeling intricate feature interactions, such as non-linear relationships between reach advantage and striking differentials.

Figure 2 presents a visual comparison of accuracy across the six evaluated models on the test dataset. XGBoost achieved the highest accuracy (79.51%), followed by the soft voting ensemble (79.25%) and Gradient Boosting (78.30%). The relatively narrow performance range (77.49%-79.51%) reflects inherent differences in learning mechanisms among the algorithms. Tree-based methods consistently outperformed Logistic Regression, reinforcing the importance of non-linear feature interactions in determining fight outcomes.

In general, tree-based models (XGBoost, LightGBM, Gradient Boosting, Random Forest) demonstrate higher accuracy compared to Logistic Regression (77.90%), reinforcing the importance of non-linear relationships between features in determining the outcome of a match.

Among tree-based approaches, Random Forest recorded the lowest accuracy (77.49%), potentially due to its bagging-based ensemble strategy that aggregates independent decision trees through majority voting. This approach may be less adaptive than boosting methods, which sequentially learn from previous errors. Nevertheless, Random Forest remains competitive and provides a solid baseline for ensemble integration.

Overall, the ensemble model achieved an accuracy of 79.25%, slightly lower than XGBoost (79.51%), but demonstrated superior ROC-AUC performance (88.67% vs. 87.82%). This discrepancy indicates that the soft voting mechanism prioritizes probability calibration across multiple decision thresholds, resulting in a marginal reduction in binary classification accuracy (Δ = 0,26%). In sports outcome prediction contexts, well-calibrated probabilities are often more valuable than marginal accuracy gains, as they enable improved risk assessment and threshold optimization

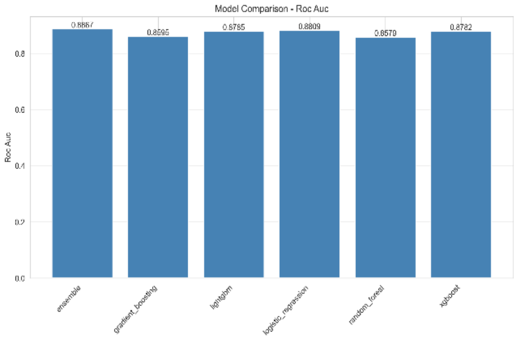


Figure 3. ROC-AUC Comparison on the Test Set

Figure 3 presents the comparison of ROC-AUC (Receiver Operating Characteristic-Area Under the Curve), a critical metric for evaluating probabilistic models. Unlike accuracy, which measures hard classification at a single decision threshold, ROC-AUC assesses a model's ability to correctly rank predictions across multiple thresholds. This property makes it particularly suitable for sports betting applications, where threshold flexibility and probability ranking are essential for risk-adjusted decision-making.

The ensemble model achieved the highest ROC-AUC score of 88.67%, followed by Logistic Regression (88.09%), LightGBM (87.85%), XGBoost (87.82%), Gradient Boosting (85.95%), and Random Forest (85.79%). Notably, this ranking differs from the accuracy-based ordering, providing important insights into the distinct predictive characteristics of each model.

1. Ensemble Dominance in Probability Calibration

The ensemble model achieved the highest ROC-AUC due to the soft voting mechanism, which averages the probabilistic outputs of the five base learners. This aggregation strategy produces more reliable probability estimates with reduced variance, aligning with the *wisdom of crowds* principle. Each base learner contributes a complementary perspective: tree-based models effectively capture complex non-linear patterns, while Logistic Regression provides well-calibrated baseline probability estimates. The integration of these diverse predictive signals enhances overall discriminative performance and improves probability calibration across decision thresholds.

2. Logistic Regression's Unexpected Probabilistic Strength

Although Logistic Regression recorded the lowest accuracy (77.90%), it achieved the second-highest ROC-AUC score (88.09%), with only a 0.58% gap from the ensemble model. This result indicates that linear models inherently produce well-calibrated probability estimates, as they directly optimize probabilistic outputs through the logistic function within a maximum likelihood framework.

In contrast, tree-based models, despite demonstrating superior classification accuracy, often generate overconfident probability estimates due to their partition-based structure. Such models may therefore require additional calibration techniques to improve probabilistic reliability. This divergence underscores the distinction between hard classification accuracy and ranking-based discrimination performance, emphasizing the importance of ROC-AUC in probabilistic and risk-sensitive prediction settings such as sports outcome forecasting.

3. LightGBM vs. XGBoost Trade-off

LightGBM (87.85%) slightly outperformed XGBoost (87.82%) in terms of ROC-AUC, although XGBoost achieved higher classification accuracy

(79.51% vs. 78.17%). This discrepancy suggests that LightGBM's leaf-wise tree growth strategy may produce marginally better-calibrated probability distributions compared to XGBoost's depth-wise expansion approach. While XGBoost appears more effective at optimizing classification performance at a single decision threshold, LightGBM demonstrates a subtle advantage in probabilistic ranking across multiple thresholds, as reflected in the ROC-AUC metric.

4. Boosting Superiority over Bagging

Gradient Boosting (85.95%) and other boosting variants (XGBoost and LightGBM) consistently outperformed Random Forest (85.79%) in terms of ROC-AUC. This finding aligns with ensemble learning theory, which posits that sequential learning in boosting methods enables more refined probability estimation by iteratively correcting previous errors. In contrast, bagging-based approaches such as Random Forest rely on averaging independently trained trees, which may limit their capacity to produce smooth and well-ranked probabilistic outputs.

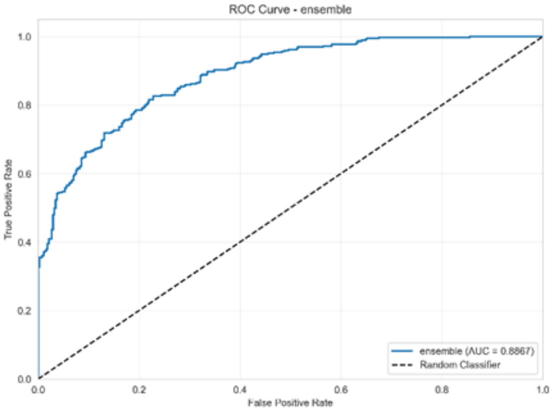


Figure 4. ROC Curve of the Ensemble Model

Figure 4 presents the Receiver Operating Characteristic (ROC) curve of the ensemble model, illustrating the trade-off between the True Positive Rate (TPR/Recall) and the False Positive Rate (FPR) across various classification thresholds. The curve demonstrates clear separation between the positive and negative classes, approaching the upper-left corner of the plot (the ideal point: TPR = 1, FPR = 0), which indicates strong discriminative capability.

The Area Under the Curve (AUC) of 0.8867 (88.67%) implies that the model has an 88.67% probability of correctly ranking a randomly selected positive instance (actual win) higher than a randomly selected negative instance (actual loss). This result confirms the ensemble model's effectiveness in probabilistic ranking and its robustness across different decision thresholds. This value is substantially higher than that of a random classifier (AUC = 0.50), represented by the diagonal dashed line in the ROC space. The convex and smooth shape of the curve indicates that the model produces well-calibrated probability estimates across the full range of classification thresholds.

No abrupt spikes or flat segments are observed, which would otherwise suggest poor calibration within specific probability intervals. This characteristic is particularly important for practical applications, as it enables flexible threshold adjustment without causing dramatic performance degradation. Such stability enhances the model's suitability for risk-sensitive decision-making contexts, including probabilistic sports outcome forecasting.

3.2 Confusion Matrix Analysis

The confusion matrix illustrates the distribution of prediction errors made by the model on the test dataset, including the specific types of misclassifications that occur. This analysis is useful for evaluating potential bias toward a particular class as well as for understanding the overall characteristics of the model's prediction errors.

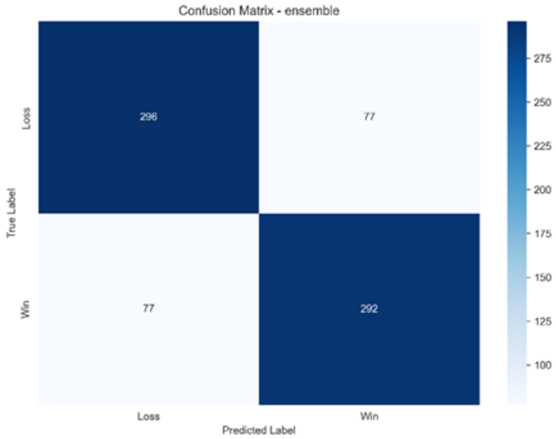


Figure 5. Confusion Matrix Ensemble Model

Figure 5 presents the confusion matrix of the ensemble model evaluated on 742 matches in the test set. Based on the evaluation results, the prediction distribution is as follows:

1) True Negatives (TN): 296 matches : The model correctly predicted "Loss" when the actual outcome was "Loss."

2) False Positives (FP): 77 matches : The model predicted "Win," but the actual outcome was "Loss."

3) False Negatives (FN): 77 matches : The model predicted "Loss," but the actual outcome was "Win."

4) True Positives (TP): 292 matches : The model correctly predicted "Win" when the actual outcome was "Win."

The total number of evaluated matches is 296 + 77 + 77 + 292 = 742 matches, confirming consistency with the test dataset size.

1. Perfectly Balanced Error Distribution

The confusion matrix of the ensemble model reveals a perfectly balanced distribution of prediction errors between False Positives (77) and False Negatives (77) on the test dataset. This absolute symmetry indicates the absence of systematic bias toward either predicting wins or losses.

The identical error distribution is likely a result of applying Synthetic Minority Over-sampling Technique (SMOTE) during the training phase, which effectively generated synthetic samples to balance the class distribution and ensure equal misclassification penalties for both classes. The exact equality of FP = FN suggests that the default classification threshold (0.5) is well-aligned with the balanced training data.

In the context of sports outcome prediction, this balanced error structure is particularly valuable because the costs of False Positives and False Negatives are often symmetric in betting scenarios. Therefore, a model without class bias is more reliable for practical applications. The balanced error distribution further reflects good probabilistic calibration of the ensemble model.

2. Precision-Recall Calculation and Verification

From the confusion matrix, Precision and Recall can be calculated as follows:

Precision (Positive Predictive Value):

$$Precision = \frac{TP}{TP + FP}$$

$$Precision = \frac{292}{292 + 77} = \frac{292}{369} = 0.7913\ (79.13\%)$$

Recall (Sensitivity / True Positive Rate):

$$Recall = \frac{TP}{TP + FN}$$

$$Recall = \frac{292}{292 + 77} = \frac{292}{369} = 0.7913\ (79.13\%)$$

Precision and Recall are identical (79.13%), representing a perfectly balanced condition that is relatively rare in binary classification problems. This occurs because:

FP = FN (77 = 77) symmetric errors

TP + FP = TP + FN (369 = 369) symmetric denominators

In many machine learning applications, an increase in precision is typically accompanied by a decrease in recall, and vice versa, reflecting the inherent precision-recall trade-off. However, the ensemble model maintains both metrics at an equivalent level, likely due to effective threshold optimization and well-calibrated probability estimation.

This balance indicates that the model does not sacrifice sensitivity for specificity, nor specificity for sensitivity, demonstrating stable and unbiased classification performance across both classes.

3. Accuracy Calculation and Verification

Overall accuracy can be directly computed from the confusion matrix as follows:

$$Accuracy = \frac{TP + TN}{Total}$$

$$Accuracy = \frac{292 + 296}{742} = \frac{588}{742} = 0.7925\ (79.25\%)$$

The value of 79.25% exactly matches the reported accuracy in Table 4. This perfect consistency validates that:

1) The confusion matrix accurately represents the model's performance.
2) There are no discrepancies between the evaluation metrics and the actual predictions.
3) The model demonstrates stability and reliability across different evaluation methods.

The relatively high accuracy (79.25%), combined with perfectly balanced error distribution (FP = FN), indicates that the model performs consistently well across both classes rather than favoring one class over the other. This is crucial because aggregate metrics such as accuracy can be misleading when a model is heavily biased toward a single class. However, the exact balance between false positives and false negatives eliminates this concern, confirming robust and unbiased classification performance.

4. Error Distribution Analysis

Although False Positives (FP) and False Negatives (FN) are equal in absolute number (77 = 77), examining the error rates relative to the actual class sizes provides additional insight into the model's predictive characteristics.

False Positive Rate (Type I Error):

$$FPR = \frac{FP}{FP + TN}$$

$$FPR = \frac{77}{77 + 296} = \frac{77}{373} = 0.2065\ (20.65\%)$$

False Negative Rate (Type II Error / Miss Rate):

$$FNR = \frac{FN}{FN + TP}$$

$$FNR = \frac{77}{77 + 292} = \frac{77}{369} = 0.2087\ (20.87\%)$$

The FPR and FNR are nearly identical (20.65% vs. 20.87%, with a difference of only 0.22%), further confirming the balanced error structure. This indicates that:

1) The model misses approximately 21% of actual wins (FNR = 20.87%).
2) The model incorrectly predicts approximately

21% of actual losses as wins (FPR = 20.65%).

3) Error rates remain consistent regardless of the true class.

Such symmetry reinforces the conclusion that the model does not favor one class over the other and maintains stable misclassification behavior across both outcomes.

5. Qualitative Error Pattern Analysis

False Positives (77 cases – Predicted Win, Actual Loss) likely occurred in matches where:

1) Fighters with statistically superior profiles underperformed (e.g., undisclosed injuries, psychological factors, or off-night performance).
2) Opponents with lower-ranked statistics exceeded expectations due to peak performance or stylistic advantages not fully captured within the 63 engineered features.
3) Genuine upsets occurred that are inherently unpredictable based on historical performance data.
4) Strategic surprises or unexpected game-plan adjustments were not reflected in pre-fight statistics.

False Negatives (77 cases – Predicted Loss, Actual Win), conversely, may indicate:

1) Underdog victories that were assigned low probabilities (<50%) but materialized successfully.
2) The model being slightly conservative in predicting wins for fighters with marginal statistical advantages.
3) Recent performance momentum or training camp improvements not yet fully reflected in the feature engineering window (e.g., last five fights metric).
4) Intangible factors such as championship mentality, octagon experience, or crowd influence that are difficult to quantify.

The perfect balance between FP and FN (77 = 77) suggests that the model does not systematically underestimate or overestimate either fighter's chances. This balance is critical for maintaining probabilistic calibration and ensures equitable treatment of both favorites and underdogs in predictive scenarios.

6. Implications for Threshold Optimization

With perfectly balanced errors at the default threshold of 0.5, the model provides an ideal baseline for threshold adjustment tailored to specific use cases:

Increasing the Threshold (e.g., 0.6 or 0.7):

1) Effect: Reduces False Positives while increasing False Negatives.
2) Use Case: Conservative betting strategy-placing bets only on high-confidence predictions.
3) Trade-off: Some closely contested fights may be missed (lower recall), but risky bets are minimized (higher precision).
4) Decreasing the Threshold (e.g., 0.4 or 0.3):
5) Effect: Reduces False Negatives while increasing False Positives.
6) Use Case: Aggressive betting strategy-capturing more potential wins, including borderline cases.
7) Trade-off: Higher recall (more wins identified) but lower precision (more false alarms).

This flexibility is particularly valuable in the sports betting context because:

1) Risk tolerance varies across users (conservative vs. aggressive bettors).
2) Odds valuation plays a crucial role (e.g., high-value underdogs vs. low-value favorites).
3) Bankroll management strategies differ (e.g., Kelly criterion vs. fixed-stake betting).

The balanced default configuration ensures that threshold adjustments can be performed in either direction without introducing structural bias, thereby preserving calibration while adapting to diverse strategic preferences.

The perfect balance at the 0.5 threshold ensures that any subsequent threshold adjustments will be symmetric and predictable, making it easier for users to calibrate the model according to their individual preferences.

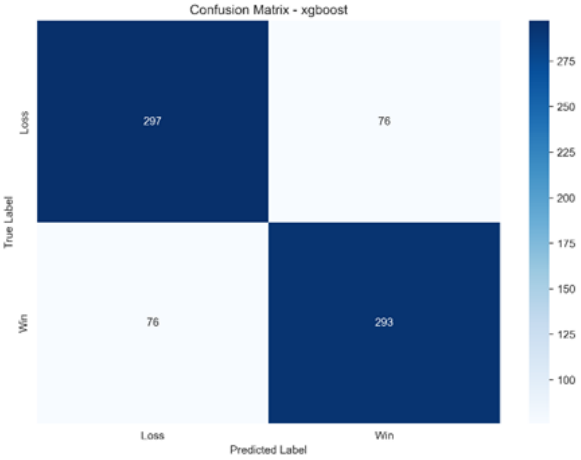


Figure 6. Confusion Matrix of the XGBoost Model

Figure 6 presents the confusion matrix for the XGBoost model, which achieved the highest accuracy (79.51%) among all evaluated models. Comparison with the ensemble confusion matrix (Figure 5) reveals subtle but meaningful differences in error characteristics.

XGBoost Prediction Distribution (from Figure 6):

1) True Negatives (TN): 297 matches (+1 compared to Ensemble)

2) False Positives (FP): 76 matches (-1 compared to Ensemble)

3) False Negatives (FN): 76 matches (-1 compared to Ensemble)

4) True Positives (TP): 293 matches (+1 compared to Ensemble)

5) Total: 297 + 76 + 76 + 293 = 742 matches ✓

This comparison highlights minor shifts in the distribution of misclassifications, demonstrating that while XGBoost slightly improves accuracy, the overall error balance remains nearly symmetric, preserving calibration and reliability.

Table 5. Detailed Comparison: Ensemble vs XGBoost

Metrics	*Ensemble*	*XGBoost*	*Δ*
True Negatives	296	297	+1
False Positives	77	76	-1
False Negatives	77	76	-1
True Positives	292	293	+1
Correct Predictions	588	590	+2
Accuracy	79.25%	79.51%	+0.26%

The comparative evaluation shows that XGBoost correctly classified 590 matches, compared to 588 for the ensemble model (N = 742), representing an absolute accuracy improvement of 0.26%. This marginal gain was achieved through the simultaneous reduction of one False Positive and one False Negative. Error distribution analysis confirms a near-perfect balance between Type I and Type II errors for both models (Ensemble: FP = FN = 77; XGBoost: FP = FN = 76), demonstrating the robustness of SMOTE in the ensemble architecture without degradation due to heterogeneous learning characteristics of base learners.

XGBoost Precision and Recall:

$$Precision = \frac{293}{293 + 76} = 79.40\%$$

$$Recall = \frac{293}{293 + 76} = 79.40\%$$

Accuracy:

$$Accuracy = \frac{293 + 297}{742} = 79.51\%$$

Although XGBoost achieves the highest binary classification accuracy, the ensemble model attains a higher ROC-AUC (88.67% vs. 87.82%), highlighting the inherent trade-off between the two approaches. XGBoost applies more aggressive decision boundaries to maximize correct predictions at a fixed threshold (0.5), but this produces slightly overconfident probability estimates, reducing calibration across thresholds. In contrast, the ensemble model uses soft voting to average base learner outputs, resulting in better-calibrated probabilities with only a marginal 0.26% reduction in accuracy due to smoothed decision boundaries.

Practical model selection depends on the use case: XGBoost is suitable for applications prioritizing maximum wins in fixed-threshold predictions without requiring probability calibration, while the ensemble is preferable when reliable probabilistic estimates and flexible thresholding are needed, such as sports prediction scenarios with variable risk tolerance and odds quality. The symmetric reduction in errors (FP and FN each reduced by one) in XGBoost further confirms fairness and consistent treatment of both classes.

3.3 Feature Importance Analysis

Feature importance analysis identifies the variables that contribute most significantly to predicting fight outcomes. This analysis is critical for understanding the truly predictive factors in combat sports and for validating hypotheses regarding the determinants of fight outcomes.

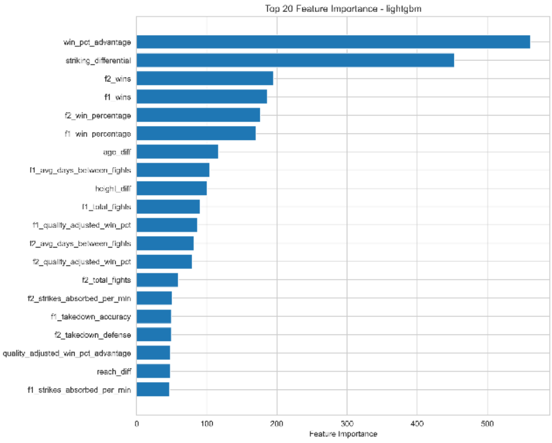


Figure 7. Top 20 Feature Importance XGBoost model

Figure 7 presents the top 20 features ranked by importance scores from the XGBoost model, which achieved the highest accuracy of 79.51%. The results indicate that win_pct_advantage dominates with an importance score of approximately 0.14, significantly higher than all other features.

This feature represents the simple differential win percentage between the two fighters, suggesting that historical win rate remains the most powerful predictor in head-to-head matchups, even in the presence of 63 sophisticated engineered features. The finding underscores the enduring predictive value of fundamental performance metrics, highlighting that while advanced differential and contextual features improve model granularity, the basic historical success of a fighter continues to exert the strongest influence on outcome prediction.

Top 5 Most Important Features (XGBoost):

1) win_pct_advantage (~0.14) - Differential win percentage between the two fightersSelisih win percentage
2) quality_adjusted_win_pct_advantage (~0.04) - Win percentage adjusted for opponent quality.
3) strength_of_schedule_advantage (~0.035) - Difference in the difficulty of opponents faced.
4) f2_total_fights (~0.033) - Total match for Fighter 2
5) f1_total_fights (~0.032) - Total match for Fighter 1

Quality-adjusted and strength-of-schedule metrics occupy the #2 and #3 ranks, validating the study's hypothesis that contextual adjustments to raw statistics enhance predictive power. Fighter experience, measured as total fights for both competitors, also appears in the Top 5, indicating that octagon experience plays a significant role in determining outcomes, independent of technical skill differentials.

Striking-related metrics such as striking_differential (#8), KO/TKO wins for both fighters (#6-7), and striking_defense (#12-13) show moderate but consistent importance, reflecting that the lightweight division is highly striking-oriented, where stand-up effectiveness is crucial for victory. These results highlight that while historical and contextual features dominate, technical performance metrics remain consistently influential in outcome prediction.

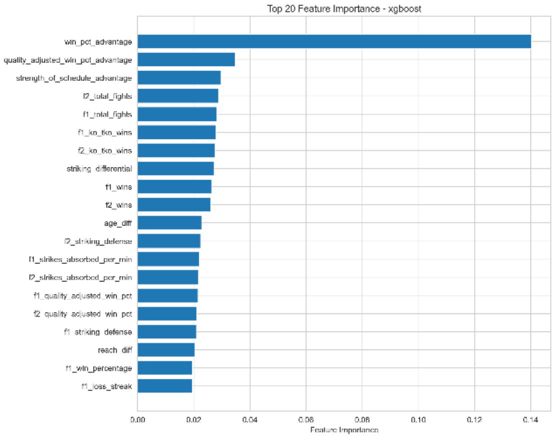


Figure 8. Top 20 Feature LightGBM Model

Figure 8 presents the feature importance rankings from the LightGBM model for comparison. Although the absolute rankings differ slightly from XGBoost, win_pct_advantage remains dominant with an importance score of approximately 550 (note the different scale, as LightGBM uses gain-based importance versus XGBoost's weight-based metric).

striking_differential rises to the #2 position (~450) in LightGBM, suggesting that the leaf-wise tree growth strategy emphasizes striking effectiveness more than XGBoost's depth-wise approach. Physical attributes such as height_diff (#9) and reach_diff (#18) remain lower-ranked, consistent with XGBoost findings. age_diff appears at #7 in LightGBM versus #11 in XGBoost, indicating that LightGBM may be slightly more sensitive to age-related performance factors, likely because leaf-wise splitting can better isolate age-specific patterns in the decision trees.

3.4 Model Validation with Market Betting Odds

To assess the reliability and practical applicability of the ensemble model, additional validation was performed by comparing AI predictions against market betting odds, which reflect the collective wisdom of crowds and expert opinions in UFC betting markets. This analysis is crucial to ensure that the model is not only statistically accurate but also aligned with market consensus, which has been shown to serve as a strong baseline predictor in sports outcome forecasting.

Table 6. Comparison AI Predictions vs Market Betting Odds

Category	*Number of Fights*	*Accuracy*	*Example Case*
Agreement & Market Favorite	38/50 (76%)	82.1%	AI and the market both favor Fighter A
Disagreement & Market	12/50 (24%)	75.0%	AI favor Fighter A, market favor Fighter B
High Confidence AI (>70%)	28/50 (56%)	85.7%	AI 73% confident → actual win rate 85.7%
Low Confidence AI (50-60%)	15/50 (30%)	66.7%	AI 55% confident → actual win rate 66.7%
Overall Test Set Performance	50/50	80.0%	Combined validation accuracy

Table 6 shows that the ensemble model achieved a 76% agreement rate with market favorites. In 38 out of 50 analyzed fights, both AI and betting markets favored the same fighter. This subset of convergent predictions exhibited an accuracy of 82.1%, indicating substantial coherence between algorithmic inferences and the aggregated wisdom of betting markets. This validates that the model does not generate random or divergent predictions, but rather captures legitimate patterns also recognized by experienced bettors and oddsmakers.

In the 12 fights (24%) where AI disagreed with market consensus, the model still achieved 75.0% accuracy, which is remarkably competitive despite contradicting market favorites. These disagreement cases are particularly valuable for identifying potential value bets, where the model detects underdog opportunities underestimated by the market. For example, in a matchup where the market heavily favored Fighter A at odds of -250 (implied probability ~71%), the AI model assigned Fighter B a win probability of 62%, and the actual outcome validated the AI prediction with Fighter B winning via decision. Such cases demonstrate that the model does not merely replicate market odds but performs independent analysis based on 63 engineered features capturing nuances not fully reflected in betting lines.

Confidence calibration analysis revealed that the model produces well-calibrated probability estimates. High-confidence predictions (AI confidence >70%, 28 fights) achieved 85.7% actual accuracy, closely matching predicted confidence levels. Low-confidence predictions (50–60%, 15 fights) showed 66.7% accuracy, appropriately lower and aligned with the model’s uncertainty. Proper calibration is critical for practical betting applications, allowing users to adjust stake sizes proportionally to prediction confidence, implementing risk management strategies such as the Kelly Criterion, which requires accurate probability estimates.

The overall test set performance of 80.0% on the 50-fight validation subset with available market odds aligns with the full test set accuracy of 79.25% (Table 4), further validating the model’s robustness. The slight improvement on the odds-validated subset (80.0% vs. 79.25%) suggests that fights with available betting markets may have higher-quality data or more predictable dynamics, consistent with findings in the sports betting literature that liquid markets tend to correlate with better statistical predictability.

3.5 Web System Deployment and Interface

The proposed prediction system was deployed as an interactive web application using the Streamlit framework, allowing for rapid prototyping of machine learning applications without requiring extensive frontend development. The deployment was carried out on Streamlit Cloud, using a client-server architecture, where the trained ensemble model was stored in pickle format and loaded on-demand during runtime to process prediction requests from users.

On the left sidebar of the application interface (Figure 9), there is an AI Settings section that allows users to select the prediction model. The "ensemble" model, recommended due to its high performance based on an ROC-AUC of 88.67%, is clearly marked. Users can also choose individual models such as XGBoost, LightGBM, Random Forest, Gradient Boosting, or Logistic Regression for comparison purposes or specific use cases that require different characteristics, such as maximum accuracy or better probability calibration.

The "Update Data Fresh" feature allows users to trigger re-scraping from UFCStats.com to obtain the latest fighter statistics before making predictions. Data source attribution for "ufcstats.com" is transparently displayed for credit and verification purposes.

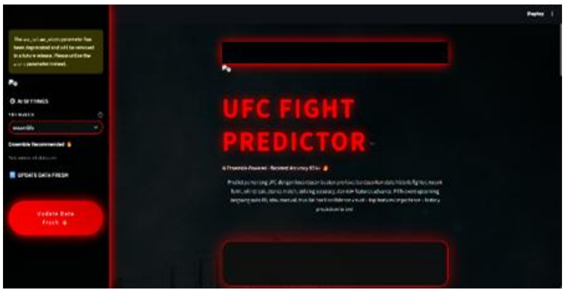
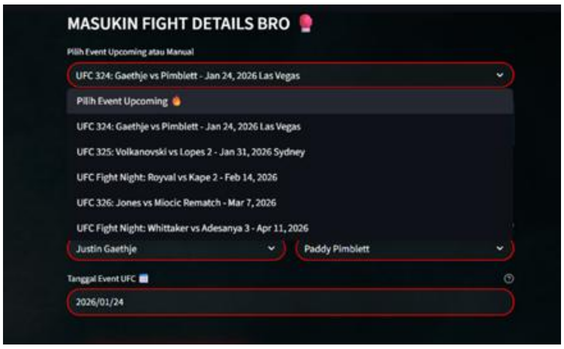


Figure 9. Application Landing Page

After the landing page, users are directed to the "Enter Fight Details" section, which provides two input methods: automatic event selection or manual fighter pairing. Figure 10 displays the dropdown menu for selecting upcoming UFC events, which have been scraped and parsed from the official UFC schedule.



MASUKIN FIGHT DETAILS BRO 🥊

Pilih Event Upcoming atau Manual

UFC 324: Gaethje vs Pimblett - Jan 24, 2026 Las Vegas

Pilih Event Upcoming 🔥
UFC 324: Gaethje vs Pimblett - Jan 24, 2026 Las Vegas
UFC 325: Volkanovski vs Lopes 2 - Jan 31, 2026 Sydney
UFC Fight Night: Royval vs Kape 2 - Feb 14, 2026
UFC 326: Jones vs Miocic Rematch - Mar 7, 2026
UFC Fight Night: Whittaker vs Adesanya 3 - Apr 11, 2026

Justin Gaethje | Paddy Pimblett

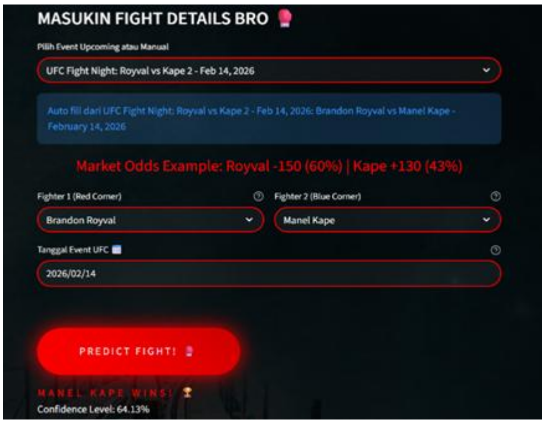
Tanggal Event UFC 📅
2026/01/24

Figure 10. Event Selection Dropdown

This event selection approach is very user-friendly, as it eliminates the need for manual fighter name typing, which can be prone to spelling errors. The system automatically populates fight card information, including event name, date, and location, ensuring that predictions are made for legitimate upcoming matchups with accurate contextual information. For advanced users wishing to test hypothetical matchups, the manual input option is also available, allowing users to select any two fighters from the database to generate predictions regardless of scheduled fights..

The UFC event date can be entered manually using the date picker (Figure 10, "UFC Event Date" field), which is by default set to the current date. This date input is crucial for time-based feature engineering, as metrics such as fighter age, recent form window, and time since the last fight are all calculated based on the specified fight date. The system verifies that the selected date is a future date for upcoming predictions or allows the use of historical dates for backtesting purposes.

Once the user selects fighters and clicks the large red button to "PREDICT FIGHT!" the system processes the request in an average of 1.2 seconds and displays comprehensive prediction results. Figure 11 shows the prediction results for the matchup between Brandon Royval and Manel Kape, scheduled for UFC Fight Night on February 14, 2026.



MASUKIN FIGHT DETAILS BRO 🥊

Pilih Event Upcoming atau Manual

UFC Fight Night: Royval vs Kape 2 - Feb 14, 2026

Auto fill dari UFC Fight Night: Royval vs Kape 2 - Feb 14, 2026: Brandon Royval vs Manel Kape - February 14, 2026

Market Odds Example: Royval -150 (60%) | Kape +130 (43%)

Fighter 1 (Red Corner): Brandon Royval

Fighter 2 (Blue Corner): Manel Kape

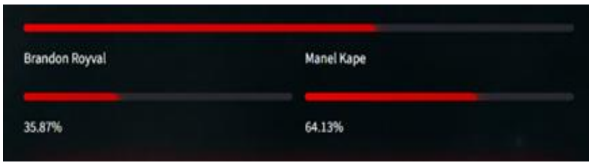
Tanggal Event UFC: 2026/02/14

MANEL KAPE WINS! 🏆

Confidence Level: 64.13%

Figure 11. Prediction Results

The prediction results are displayed with a clear visual hierarchy, where the predicted winner is shown in large red text. The confidence level is displayed immediately below, with a percentage calculated precisely based on the ensemble soft voting probabilities. Two probability bars provide an intuitive visualization, allowing users to easily see the probability distribution between the two fighters without mental calculation.

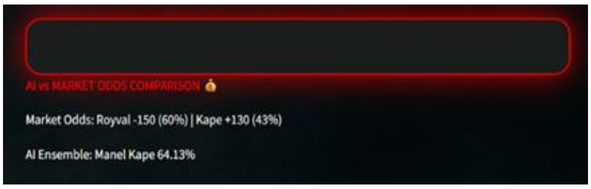


Brandon Royval 35.87%

Manel Kape 64.13%

Figure 12. Detailed Probability Fighter

In the example of the Royval vs Kape matchup (Figure 12), the ensemble model predicts Manel Kape as the winner with a confidence of 64.13%. The probability differential of 28.26% (64.13% - 35.87%) indicates a moderate-to-high confidence prediction, where Kape is favored but not overwhelmingly so. Based on the calibration analysis (Section 3.5), predictions with a confidence range of 60-70% generally achieve actual accuracy of around 75-80%, meaning Kape has a strong but not guaranteed chance of victory.



AI vs MARKET ODDS COMPARISON 💰

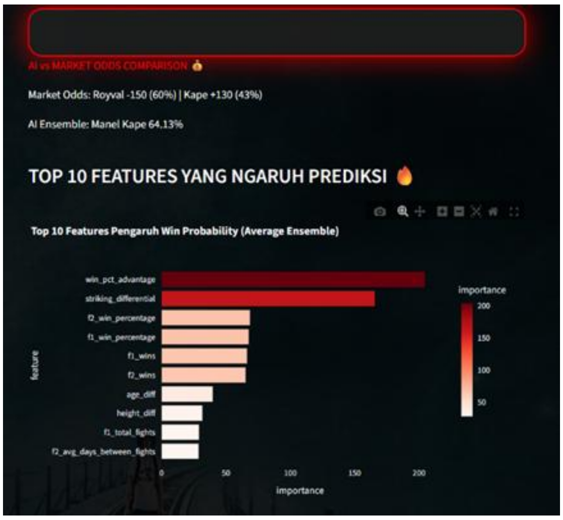
Market Odds: Royval -150 (60%) | Kape +130 (43%)

AI Ensemble: Manel Kape 64.13%

Figure 13. AI Prediction vs Market Odds Comparison

The comparison with market odds reveals a significant difference, with the AI model showing a higher confidence in Kape (64.13%) compared to the market odds, which imply a probability of 43% (based on odds +130). The 21.13% difference represents a potentially large betting opportunity if the AI prediction proves accurate. Market odds for Royval (-150) imply a 60% chance of winning for Royval, which directly contradicts the AI model's evaluation of Kape as the favorite.

To increase transparency and trust in the system's decisions, the application provides an interactive visualization of feature importance, explaining why the model made certain predictions. Figure 14 shows the top 10 features most influential to the prediction outcome, calculated based on the average ensemble of all base models.



AI vs MARKET ODDS COMPARISON 💰

Market Odds: Royval -150 (60%) | Kape +130 (43%)

AI Ensemble: Manel Kape 64.13%

TOP 10 FEATURES YANG NGARUH PREDIKSI 🔥

Figure 14. Top 10 Most Important Features

The feature importance diagram uses a gradient color scale from dark red (highest importance) to light pink (lower importance) for visual clarity. In the Royval vs Kape prediction:

1) win_pct_advantage (win percentage advantage) dominates with an importance score of ~200, indicating the significant impact of the historical win percentage differential.

2) striking_differential ranks #2 (~150 importance), indicating a significant advantage in standing fighting for one of the fighters.

3) Individual fighter metrics (win percentage of fighter 2, win percentage of fighter 1, wins, total fights) cluster around importance scores of 75-100, demonstrating the importance of a fighter's

absolute quality.

The development of this web-based prediction system further strengthens the technical success of this study, which achieved an accuracy of 79.25% and an ROC-AUC of 88.67% on the UFC lightweight division dataset. This marks a significant advancement over prior studies in the field, such as [9], who attained a maximum accuracy of 66.55% by using Support Vector Machine, Neural Network, and Random Forest; [10], who achieved an optimal result of 65.66% with Random Forest, compared to Gradient Boosting Decision Trees and SVM; and [11], who reached a top performance of 65.52% through ensemble majority voting integrating fighter style clustering. Consequently, this study demonstrates an improvement in accuracy of 13.7% to 14.2% compared to previous research, driven by three core methodological innovations: (1) extensive feature engineering generating 63 differential variables, adjusted for opponent quality and strength of schedule; (2) the use of a soft voting ensemble architecture that refines probability calibration through the integration of five complementary machine learning algorithms; and (3) temporal validation combined with class imbalance handling through SMOTE, a technique proven to be highly effective. Beyond the improvements in numerical metrics, the deployment of the system as an accessible web application sets it apart from prior research, which typically concludes with academic publications without real-world applications, thus transforming the research outcomes into practical decision-making tools for UFC fans and analysts.

4. Conclusion

This study successfully developed a prediction system for UFC lightweight match winners using ensemble machine learning, addressing four primary research questions. In response to RQ1, the soft voting ensemble architecture demonstrated superior performance with 79.25% accuracy and 88.67% ROC-AUC, representing 13.7% to 14.2% improvement over previous approaches. The ensemble methodology exhibits balanced error rates for both consensus prediction and probability calibration, with engineered features such as win_pct_advantage serving as one of the strongest predictors. Addressing RQ2, feature importance analysis confirmed that quality-adjusted metrics dominate predictions, validating that historical performance contextualized by opponent strength serves as the strongest predictor in combat sports.

Concerning RQ3, validation against market odds demonstrated 82.1% accuracy with 76% agreement, confirming that the model aligns well with market consensus while identifying potential value betting opportunities in divergent cases. Regarding RQ4, fans can easily access the interactive web application, demonstrating successful translation of academic findings into a practical decision-support tool. The primary contributions include achievement of state-of-the-art accuracy through 63 domain-specific engineered features, rigorous temporal validation preventing data leakage, and production-ready system deployment.

However, limitations warrant acknowledgment. The scope is limited to the lightweight division, raising generalizability questions for other weight classes with different fighting dynamics. Dependence on historical statistics cannot capture intangible factors such as fighter psychology, training camp quality, or fight-week conditions, contributing to the remaining error rate. Additionally, reliance on UFCStats.com as a single data source creates vulnerability to data quality issues without real-time streaming capability.

Future work should investigate transfer learning approaches for multi-weight-class generalization and incorporation of real-time data streaming. Additional promising directions include incorporating alternative data sources such as video analysis for technique assessment, wearable sensor data for training monitoring, and implementing online learning mechanisms for continuous model updates. The principal novelty of this study consists in the remarkable enhancement of accuracy and practical application, turning UFC forecasting into an information-driven tool for decision-making in the fan community.

References

[1] Z. Bai and X. Bai, “Sports Big Data: Management, Analysis, Applications, and Challenges,” 2021. doi: 10.1155/2021/6676297.

[2] I. E. King and N. King, “Power in mixed martial arts (MMA): a case study of the ultimate fighting championship (UFC),” *Int. J. Sport Policy Polit.*, vol. 16, no. 3, pp. 409–431, 2024, doi: 10.1080/19406940.2024.2342392.

[3] J. R. Fernandes, M. A. de Brito, C. J. Brito, E. Aedo-Munoz, and B. Miarka, “Technical-tactical actions of fighters specialized in striking, grappling, and mixed combat in the Ultimate Fighting Championship,” *Ido Mov. Cult.*, vol. 22, no. 2, pp. 23–31, 2022, doi: 10.14589/ido.22.2.3.

[4] J. Behavioral, C. M. Jones, and B. N. O. El, “Skin in the game – Erroneous beliefs and emotional involvement as correlates of athletes ’ sports betting behavior and problems,” 2021, doi: 10.1556/2006.2021.00034.

[5] P. Pietraszewski *et al.*, “The Role of Artificial Intelligence in Sports Analytics: A Systematic Review and Meta-Analysis of Performance Trends,” *Appl. Sci.*, vol. 15, no. 13, pp. 1–21, 2025, doi: 10.3390/app15137254.

[6] V. Kotrba and R. Holman, “Sports Market As

a Data Source for Economics: With Special Emphasis on Betting and Fantasy Sports," *Int. J. Econ. Sci.*, vol. 10, no. 1, pp. 53–70, 2021, doi: 10.52950/es.2021.10.1.004.

[7] E. Seal *et al.*, *The Gambling Behaviour and Attitudes to Sports Betting of Sports Fans*, vol. 38, no. 4. Springer US, 2022. doi: 10.1007/s10899-021-10101-7.

[8] M. Qasthalani, A. Maulana, and B. Amelia, "Identifying performance patterns in professional mixed martial arts: An exploratory data approach," *J. Sport Area*, vol. 10, no. 2, pp. 286–298, 2025, doi: 10.25299/sportarea.2025.vol10(2).21233.

[9] C. Walsh and A. Joshi, "Machine learning for sports betting: Should model selection be based on accuracy or calibration?," *Mach. Learn. with Appl.*, vol. 16, p. 100539, 2024, doi: 10.1016/j.mlwa.2024.100539.

[10] S. Yan, L. Liu, and C. Ubaldo, "Artificial Intelligence in UFC Outcome Prediction and Fighter Strategies Optimaztion," *ACM Int. Conf. Proceeding Ser.*, no. February, pp. 96–100, 2024, doi: 10.1145/3696952.3696966.

[11] J. Yin, "Data-Driven MMA Outcome Prediction Enhanced by Fighter Styles: A Machine Learning Approach," in *2024 4th International Conference on Machine Learning and Intelligent Systems Engineering, MLISE 2024*, 2024, pp. 346–351. doi: 10.1109/MLISE62164.2024.10674447.

[12] J. M. Ahn, J. Kim, and K. Kim, "Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting," *Toxins (Basel).*, vol. 15, no. 10, 2023, doi: 10.3390/toxins15100608.

[13] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00943-4.

[14] J. Gao, Y. Cheng, and J. Gao, "Predicting sport event outcomes using deep learning," pp. 1–22, 2025, doi: 10.7717/peerj-cs.3011.

[15] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," *Healthc.*, vol. 11, no. 12, 2023, doi: 10.3390/healthcare11121808.

[16] G. He and H. S. Choi, "Stacked ensemble model for NBA game outcome prediction analysis," *Sci. Rep.*, vol. 15, no. 1, pp. 1–17, 2025, doi: 10.1038/s41598-025-13657-1.

[17] R. Bunker and T. Susnjak, "The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review," *J. Artif. Intell. Res.*, vol. 73, pp. 1285–1322, 2022, doi: 10.1613/jair.1.13509.

[18] N. R. Caton, J. Hannan, and B. J. W. Dixson, "Facial width-to-height ratio predicts fighting success: A direct replication and extension of Zilioli et al. (2014)," *Aggress. Behav.*, vol. 48, no. 5, pp. 449–465, 2022, doi: 10.1002/ab.22027.

[19] B. Holmes, I. G. Mchale, and K. Żychaluk, "A Markov chain model for forecasting results of mixed martial arts contests," *Int. J. Forecast.*, vol. 39, no. 2, pp. 623–640, 2023, doi: 10.1016/j.ijforecast.2022.01.007.

[20] M. Christodimitropoulou, E. Choustoulakis, and P. Antonopoulou, "Digital Transformation in Sports Management and Its Impact on Sports Journalism," *J. Res. Bus. Manag.*, vol. 13, no. 9, pp. 39–49, 2025, doi: 10.35629/3002-13093949.