Predicting Tennis Match Upsets Using Machine Learning Approaches

Manuel Rodriguez Berdud
Department of Data Science
Florida Polytechnic University
Email: mrodriguezberdud318@floridapoly.edu

***Abstract*—Unexpected match outcomes, commonly referred to as upsets, are among the most compelling aspects of professional tennis. While player rankings are often assumed to be reliable predictors of match results, in practice, lower-ranked players frequently defeat higher-ranked opponents. This project proposes to study the conditions under which such upsets occur using publicly available ATP match data compiled by Jeff Sackmann. The dataset contains detailed match records, player rankings, surface types, and performance statistics, providing a rich basis for analysis. The focus of the study will be on identifying which contextual and performance-related factors are most strongly associated with upsets, and on exploring whether such outcomes can be predicted with reasonable accuracy.**

The project will begin with data cleaning and exploratory analysis to uncover descriptive trends related to player rankings, surfaces, and tournament settings. Predictive modeling will then be applied in order to evaluate the extent to which upsets can be anticipated from pre-match conditions. The expected contribution is twofold: first, to provide a clearer understanding of the key drivers behind upset matches; and second, to highlight the potential of computational data analysis as a tool for gaining insights into the dynamics of professional tennis. The broader significance of this work lies in its potential applications to coaching, strategy development, sports commentary, and fan engagement, where understanding unpredictability plays a central role.

***Index Terms*—Sports Analytics, Machine Learning, Classification, Tennis, Upset Prediction**

I. Introduction

Professional tennis is a sport that thrives on competition, skill, and strategy, but it also carries a unique element of uncertainty. Despite the predictive power of official ATP rankings, tournament seeding, and historical win–loss records, unexpected outcomes occur frequently. These so-called upsets, where a lower-ranked player defeats a higher-ranked opponent, often capture global attention and are remembered as defining moments of tournaments. For players, coaches, and analysts, understanding why such upsets occur is not only intellectually intriguing but also practically valuable.

Tennis upsets challenge the assumption that rankings alone can explain competitive performance. They highlight the complex interaction of factors such as playing surface, player fatigue, head-to-head history, tournament level, and even psychological elements like momentum and confidence (Fig 1). For example, a clay-court specialist ranked outside the top 50 may have a significant advantage against a top-10 opponent on clay, despite a large gap in ranking points. Similarly, players returning from injury or facing long travel schedules may underperform relative to expectations. These scenarios underscore the need for a more nuanced, data-driven approach to analyzing match outcomes.

The availability of large-scale, high-quality tennis data provides an unprecedented opportunity to study these dynamics systematically. Publicly available datasets, such as Jeff Sackmann's ATP match records, contain rich information on match results, player attributes, rankings, surfaces, and performance metrics. By leveraging computational data analysis techniques, it becomes possible to uncover hidden patterns that explain when and why upsets are most likely to occur. Beyond academic curiosity, these insights have broader implications: they can inform coaching strategies, enhance predictive modeling for broadcasters and sports analysts, and deepen fans' appreciation of the sport.

Contributions. This study makes two main contributions. First, it proposes a framework dedicated to predicting tennis upsets, a phenomenon often overlooked in prior research that mostly targets overall match outcomes. By combining contextual factors (surface, tournament level), performance statistics (serve, return, and break point efficiency), and indicators of player momentum, the project aims to identify which conditions most strongly contribute to unexpected results. Second, it compares multiple machine learning models to evaluate how well they can anticipate these upsets, offering practical insights for analysts, coaches, and fans interested in understanding unpredictability in tennis.

Expected Challenges. Two main challenges are anticipated in this project: **class imbalance** and **the definition of an upset**. Defining what qualifies as an upset presents a conceptual challenge. While a ranking-based definition is straightforward, it may overlook important contextual factors such as recent performance, surface specialization, or player fatigue. To ensure a more reliable target variable, the project will compare ranking-based and betting-odds-based definitions, aiming to capture both competitive expectations and real-world perceptions of match difficulty. Furthermore, since true upsets are relatively uncommon, the dataset is expected to be imbalanced. This imbalance can bias models toward predicting the majority class (non-upset outcomes), reducing their ability to detect rare but meaningful events. To mitigate this issue, the study will explore resampling methods such as

oversampling and class-weight adjustments, while evaluation metrics like F1-score and recall will be prioritized over simple accuracy.

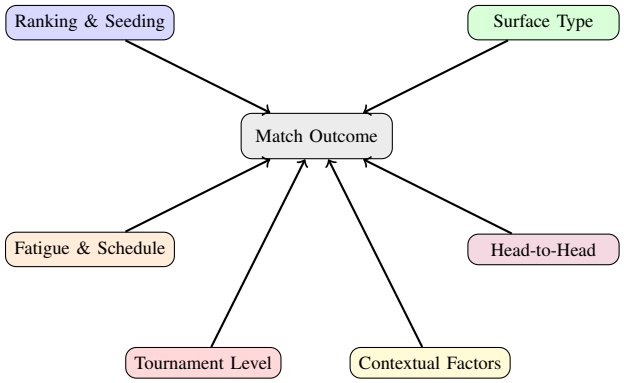

Fig. 1. Factors influencing tennis match outcomes.

II. Related Work

The prediction of tennis match outcomes, particularly upsets where lower-ranked players defeat higher-ranked opponents, has garnered significant attention in sports analytics. Machine learning (ML) methodologies have been extensively explored to forecast match results, leveraging diverse datasets and modeling techniques.

A. Machine Learning Models for Tennis Match Prediction

Early studies in tennis match prediction primarily employed traditional ML algorithms. For instance, De Seranno [1] utilized logistic regression and neural networks to predict ATP singles match outcomes, incorporating features such as player statistics and match context. Their models outperformed baseline ATP rankings, achieving notable accuracy and profitability in betting scenarios. Similarly, Dryja [2] focused on Grand Slam matches, applying decision trees, logistic regression, random forests, and XGBoost to predict match outcomes. Their findings indicated that XGBoost provided superior performance, particularly in matches with clear favorites. Gao and Kowalczyk [3] identified serve strength as a critical predictor of match outcomes using random forest models, highlighting the importance of specific technical metrics in forecasting results.

B. Incorporating Psychological and Strategic Momentum

Recent research has delved into psychological and strategic factors influencing match outcomes. Bai et al. [4] proposed a model integrating strategic and psychological momentum to predict match results in best-of-three tennis contests. Their analysis of over 66,000 matches demonstrated that both momentum types significantly impacted outcomes. Additionally, Zhai and Wang [5] introduced a Lasso-Ridge-based XGBoost model to quantify momentum effects, achieving high accuracy and emphasizing the role of momentum dynamics and game fluctuation in match predictions.

C. Real-Time and In-Match Prediction Models

Advancements in real-time data collection have facilitated in-match prediction models. Rui et al. [6] proposed a supervised ML approach to predict the flow of points in tennis matches, enhancing the granularity of match outcome predictions. Moreover, AI-based methods have been applied to utilize real-time point-by-point data to model psychological momentum during matches [7].

D. Betting Market Analysis and Upset Detection

The intersection of ML models and betting markets has been a focal point for upset detection. Rosenfield [8] developed a Python script employing logistic regression to predict match outcomes based on players' previous performances, aiming to identify potential upsets. Additionally, Zhai and Wang [5] explored meta-learning using MAML to transfer models for predicting outcomes in other sports such as ping-pong, demonstrating adaptability across domains.

E. Feature Engineering and Model Interpretability

Feature selection and model interpretability remain critical in ML-based sports predictions. Key features such as serve strength, break points, and unforced errors have been identified as significant predictors of match outcomes [3], [6]. Ensemble methods, such as soft voting, have also been effective in reducing individual model biases and errors, enhancing overall prediction accuracy [7].

Despite significant progress, existing studies often focus on predicting overall match outcomes or rely heavily on high-level player statistics, with limited attention to the specific phenomenon of upsets. Few studies systematically integrate pre-match factors, player momentum, and historical upset patterns to explicitly forecast matches where lower-ranked players defeat favorites. This gap highlights the need for specialized ML models targeting tennis upsets, which can provide more actionable insights for coaching, strategy development, and betting markets. The proposed project aims to address this gap by combining comprehensive player metrics, historical performance data, and advanced ML algorithms to improve upset prediction accuracy.

III. Data and Preprocessing

A. Data Sources

The primary dataset used in this study is the publicly available ATP match data compiled by Jeff Sackmann in GitHub [9]. This repository provided 27 CSV files that covered every ATP match played between 1998 and 2024, each of which corresponded to a single season. Fig 2 shows how the variables are grouped, which include information such as:

- Match outcomes and player identifiers
- ATP rankings and ranking points at the time of the match
- Tournament characteristics (e.g., Grand Slam, Masters, 500/250 series, Challenger)
- Surface types (hard, clay, grass, carpet)

Group	Columns	
Tournament context	tourney_id tourney_name tourney_level tourney_date best_of round	surface draw_size match_num minutes score
Player demographics and ranking	_id _seed _entry _name _hand	_ht _ioc _age _rank _rank_points
Point-aggregate performance	_ace _df _svpt _1stIn _1stWon	_2ndWon _SvGms _bpSaved _bpFaced

Fig. 2. Variable groups from the Sackmann's dataset

- Performance statistics (first serve in, aces, double faults, break points converted/saved, winners, unforced errors)
- Contextual factors (match round, player age, height, handedness, head-to-head record)

These variables provide a rich basis for exploring both contextual and performance-related factors that contribute to match upsets. Following previous studies [1]–[4], [7], the most recent CSV files are analyzed, which contain matches from 2016 to 2024.

B. Data Cleaning

The raw dataset consisted of ATP matches from 2016 to 2024 (excluding 2020 due to COVID-19), initially stored in yearly CSV files. After extensive cleaning and preprocessing, the dataset was consolidated into a consistent and reliable format. The main steps are summarized as follows:

- **Initial Assessment:** Conducted a missing value analysis across all variables. Four features with more than 50% missing values were removed due to limited analytical value.
- **Incomplete Matches:** Removed matches lacking reliable outcomes:
  - Walkovers (`RET` in score): 611 matches
  - Matches under 30 minutes: 122 matches

  In total, 1,750 matches (7.52%) were removed, leaving 21,517 valid matches.
- **Data Type Standardization:** Dates converted to `datetime`, numerical statistics properly typed, and categorical fields normalized.
- **String Standardization:** Tournament names cleaned of whitespace, surfaces mapped to four categories (Hard, Clay, Grass, Carpet), and player names standardized in capitalization and spacing.
- **Match Duration Imputation:** For 4.74% of matches with missing duration, a KNN-based imputation was applied using variables such as number of sets, total games, service statistics, and break point data. The imputed durations preserved natural distribution patterns.

Final Dataset: The cleaned dataset contains 21,517 matches across 643 tournaments and 1,074 players (original dataset had 23,267 matches). All critical match information is complete, categorical variables standardized, and data types consistent across features.

C. Target Variable Definition

In this study, two alternative definitions of an *upset* will be considered:

[Ranking-Based Upset] A match is defined as an upset if the player with the higher official ATP ranking (numerically smaller rank value) at the time of the match loses to the player with the lower ranking (numerically larger rank value). Formally, let r_A and r_B be the ATP rankings of players A and B respectively. If $r_A < r_B$ and player A loses the match, then the outcome is considered an upset.

[Betting-Odds-Based Upset] A match is defined as an upset if the player with the lower implied probability of winning, based on pre-match betting odds, wins the match. Let O_A and O_B denote the decimal odds assigned to players A and B, and let $P_A = \frac{1}{O_A}$ and $P_B = \frac{1}{O_B}$ denote their implied winning probabilities. If $P_A > P_B$ and player A loses (i.e., the underdog wins), the outcome is classified as an upset.

Initially, our study will focus on the analysis of upsets using the Ranking-based definitions. Further studies will also consider the Betting-odds-based definition.

D. Exploratory Data Analysis

An **upset** has been defined as a match where the winner's ATP ranking is at least 20 positions worse (numerically higher) than the opponent's at the time of play. This threshold filters out trivial rank differences and focuses on meaningful surprises.

Across 21,517 matches (2016–2024), 5,181 (24.1%) were classified as upsets, showing that roughly one in four matches ends with the lower-ranked player winning. While higher-ranked players dominate overall, the upset rate has increased from 20.8% in 2016 to 25.9% in 2024, suggesting growing parity in men's tennis.

Furthermore, to better understand the structure of the dataset and the behavior of upsets in professional tennis, we want to look at a few additional descriptive patterns before modeling. In particular, we examine how upset frequency varies across surfaces and tournament levels:

*a) **Upset Rate by Surface**:* Surface type influences match volatility, player performance, and physical demands. We compute the upset rate separately for hard, clay, and grass events. This helps identify whether certain surfaces exhibit higher unpredictability.

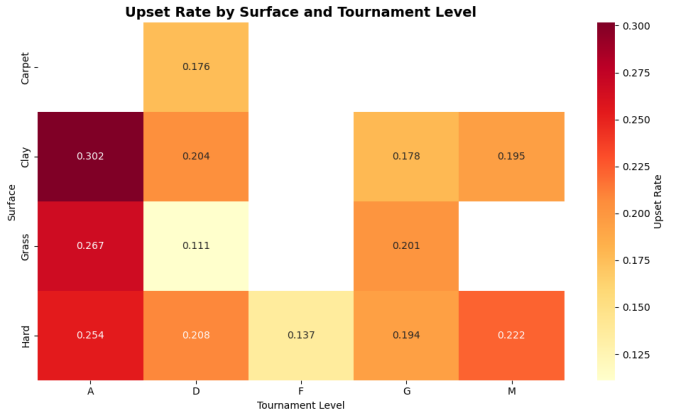

Fig. 3. Upset rate by surface and tournament level. A - ATP, D - Davis Cup, F - Finals Cup, G - Grand Slam, M - Masters 1000

*b) **Upset Rate by Tournament Level**:* We also analyze how upset likelihood changes with the tournament category (Grand Slam, Masters 1000, ATP 500, ATP 250, Davis Cup...). Higher-tier tournaments often exhibit more consistent performance among top players, which may result in fewer upsets.

Fig 3 shows how upset frequency varies noticeably across surfaces and tournament levels. Clay courts exhibit the highest upset rates, particularly in level A events (ATP 250 and ATP 500). Hard courts display more balanced rates, while carpet events show moderate volatility. Overall, both surface type and tournament category influence how likely unexpected defeats are to occur.

E. Feature Engineering

To capture the multifaceted nature of tennis upsets, raw statistics were transformed into explanatory features across five main categories, as shown in Fig 4. Each of these addresses limitations of relying solely on ATP rankings.

Serve Aggressiveness: Serve-related features such as ace-to-double-fault ratios, first-serve efficiency, and break point conversion provide insight into a player's balance between power and consistency, long recognized as decisive factors in match outcomes [3].

Recent Form: Short-term performance indicators reflect momentum and fitness more accurately than long-term records. Including rolling form metrics enables the model to adapt to a player's current competitive state [2].

Surface Specialization: Since surface type strongly conditions play style and effectiveness, features such as surface-specific win rates and adaptation measures account for contextual variability across clay, grass, and hard courts [1].

Head-to-Head History: Historical records between two players capture psychological and tactical familiarity. These features, particularly when surface-specific, complement form metrics and contextualize player match-ups.

Psychological Momentum: Metrics such as break point performance, set momentum, and winning streaks capture confidence and resilience, factors shown to influence outcomes beyond technical ability [11].

By combining these categories, the feature set extends beyond rankings to include dynamic, contextual, and psychological dimensions, providing a more comprehensive foundation for upset prediction.

F. Dataset Partitioning

To prevent data leakage, we use a **time-aware split** rather than random partitioning [1]. Training is performed on earlier seasons (2016–2022), while testing uses later seasons (2023–2024), mimicking real-world forecasting. For hyperparameter tuning, we apply **rolling origin cross-validation**, where models are trained on past matches and validated on the next chronological segment [2]. This approach ensures realistic evaluation and avoids using future information in training.

IV. Methodology

This study employs a comparative machine learning framework to predict tennis match upsets. The methodology is designed to ensure reproducibility, fairness in model evaluation, and robustness against the challenges posed by class imbalance and temporal dependencies in the data.

A. Model Selection

This study applies three supervised learning algorithms: Logistic Regression, Random Forest, and XGBoost. Each of these models brings different strengths in terms of interpretability, computational efficiency, and ability to capture complex relationships in the data. By comparing them, we aim to balance predictive performance with interpretability and robustness [12].

*1) **Logistic Regression**:* Logistic Regression (LR) is a baseline classification model that estimates the probability of an outcome using a linear decision boundary. Its primary advantage lies in interpretability: coefficients can be directly linked to the contribution of each feature toward the likelihood of an upset. LR is computationally efficient and resistant to overfitting in high-bias, low-variance contexts. However, its performance may be limited in capturing nonlinear relationships and feature interactions common in sports data. We expect Logistic Regression to provide solid baseline accuracy and well-calibrated probabilities, but lower predictive power compared to ensemble methods.

*2) **Random Forest**:* Random Forest (RF) is an ensemble learning algorithm that builds multiple decision trees on bootstrapped subsets of the data and aggregates their predictions. Its strength lies in capturing nonlinear interactions between features and handling complex, high-dimensional data. Random Forest is less sensitive to overfitting than individual decision trees and provides interpretable feature importance scores, which are valuable for identifying key predictors of upsets. We expect RF to outperform Logistic Regression in predictive accuracy due to its flexibility, while still maintaining robustness across different subsets of the data [13].

	Category	Engineered Feature	Source Variables
0	Serve	ace_df_ratio_diff	w_ace, w_df, l_ace, l_df
1	Serve	first_serve_pct_diff	w_1stIn, w_svpt, l_1stIn, l_svpt
2	Serve	first_serve_won_pct_diff	w_1stWon, w_1stIn, l_1stWon, l_1stIn
3	Serve	second_serve_won_pct_diff	w_2ndWon, l_2ndWon
4	Serve	serve_efficiency_diff	w_SvGms, l_SvGms
5	Form	form_3m_diff, form_6m_diff, form_12m_diff	winner_rank, loser_rank, tourney_date
6	Surface	surface_winrate_diff	surface, winner_name, loser_name
7	Surface	surface_adaptation_diff	surface, winner_name, loser_name
8	H2H	h2h_record_diff	winner_name, loser_name
9	H2H	h2h_surface_diff	winner_name, loser_name, surface
10	Momentum	set_momentum	score
11	Momentum	streak_diff	winner_name, loser_name
12	Momentum	tournament_stage	round

Fig. 4. Summary of the engineered features from the original variables in the dataset

*3) **XGBoost**:* Extreme Gradient Boosting (XGBoost) is a gradient boosting framework optimized for speed and performance. It builds trees sequentially, with each new tree correcting errors made by the previous ones. XGBoost incorporates regularization, efficient handling of missing values, and parallel processing, making it particularly effective on structured datasets such as tennis match statistics. Previous studies in sports analytics have shown that boosting algorithms frequently achieve state-of-the-art performance in predictive tasks. We therefore expect XGBoost to deliver the highest accuracy and discrimination power among the models considered, though at the cost of reduced interpretability relative to Logistic Regression and Random Forest [14].

Neural network models were considered as a potential extension to this study [15]. Since our dataset consists of tabular, feature-engineered inputs for which tree-based models are typically better suited and more sample-efficient, NN models weren´t applied in the study. Furthermore, neural networks would require substantially more data, additional hyperparameter tuning, and more complex regularization to avoid overfitting, while offering limited expected performance gains in this setting. For these reasons, we focus on classical and ensemble methods that align more naturally with the structure and scale of the available data.

B. Class Imbalance Approach

Since upsets are less likely than non-upsets, the dataset is highly imbalanced, which can lead models to favor majority predictions. To reduce this bias, class weights are adjusted so that misclassifying an upset incurs a higher penalty during training. This weighting approach is supported by Logistic Regression, Random Forest, and XGBoost. In addition, the Synthetic Minority Oversampling Technique (SMOTE) is tested to create synthetic upset examples and improve representation of the minority class (results can be seen in Fig 6, where both recall and F1-score show improvement)

C. Model Interpretability

Understanding why a model predicts an upset is as important as the prediction itself. Feature importance is analyzed

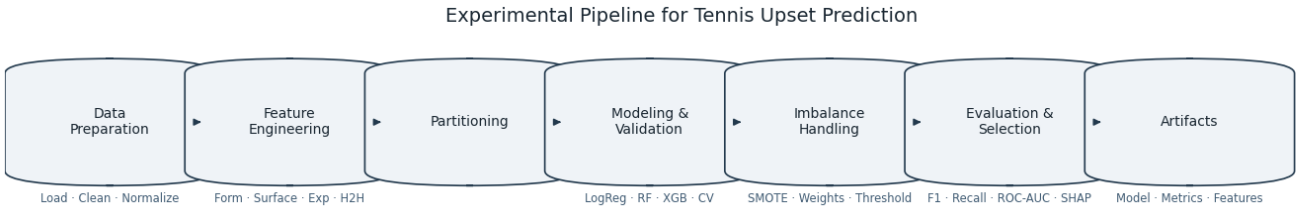

Fig. 5. Experimental pipeline for tennis upset prediction

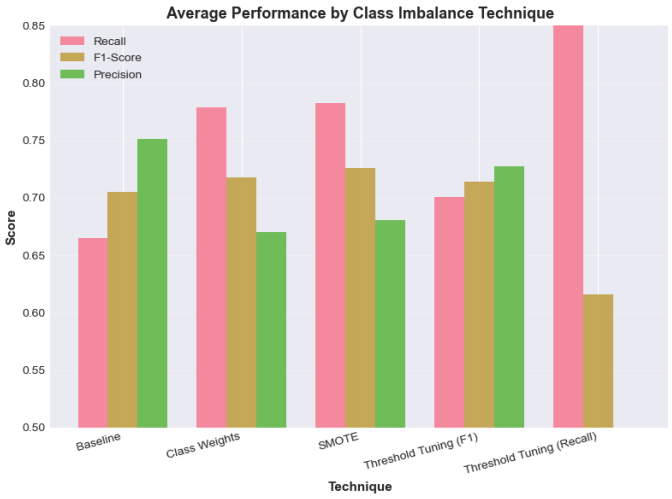

Fig. 6. Average performance by class imbalance technique, including Class Weights, SMOTE, and Threshold Tuning

for Random Forest and XGBoost to identify the factors that contribute most to model decisions. Two methods are used: (1) impurity-based importance, which measures how much each feature reduces prediction error, and (2) SHAP values, which show each feature's contribution to individual predictions. These analyses highlight which contextual and performance variables—such as surface type, recent form, or serve effectiveness—most influence upset outcomes.

D. Evaluation Metrics

Given that upsets represent a minority class in the dataset, model performance is evaluated using metrics that account for class imbalance and the need for reliable upset detection. The following metrics are used:

- **Recall (Sensitivity)** – Measures the proportion of actual upsets correctly identified by the model. High recall ensures that most true upsets are detected, making it one of the key metrics for this study.
- **F1-Score** – The harmonic mean of precision and recall, providing a single measure that balances both false positives and false negatives. Because upset matches are less probable, the F1-score offers a more informative assessment of overall performance than accuracy alone.
- **Confusion Matrix** – Summarizes the model's classification results by displaying counts of true positives (correct upset predictions), true negatives (correct non-upset predictions), false positives, and false negatives. This allows detailed analysis of model behavior and identification of systematic misclassifications.
- **ROC AUC** – Evaluates the model's ability to discriminate between upset and non-upset outcomes over all thresholds.

V. Experiment & Results

After explaining how tennis upsets are predicted, it is time to apply the theory to our cleaned and engineered datasets. Before modeling, features that were highly correlated with each other were removed to reduce **multicollinearity** and stabilize model coefficients. Furthermore, **standardizing features** helped the machine learning models classify relevant features better.

A. Modeling Pipeline

To predict tennis match upsets, we designed a structured data modeling workflow that spanned from raw ATP match data (2016–2024) to calibrated predictive models. After data cleaning, imputation, and outlier detection, we **engineered 83 performance and contextual features** capturing player form, surface specialization, head-to-head history, and experience gaps.

For model training, we employed a time-aware split, ensuring chronological integrity between training and test sets. Three primary algorithms were evaluated: *Logistic Regression, Random Forest, and XGBoost*. Each was tested under baseline and imbalance-corrected settings using **nested cross-validation** for parameter tuning and performance stability.

An experimental workflow is shown in Fig 5, where each technical step in the study is represented from left to right.

B. Baseline Models

Baseline modeling established initial predictive benchmarks using Logistic Regression and Random Forests trained on the engineered feature set without imbalance correction.

Logistic Regression achieved an **F1-score of 0.683 and ROC-AUC of 0.896**, serving as an interpretable linear baseline.

Random Forest improved slightly with an **F1-score of 0.688 and ROC-AUC of 0.914**. This classification technique initially set some of the most important features for predicting

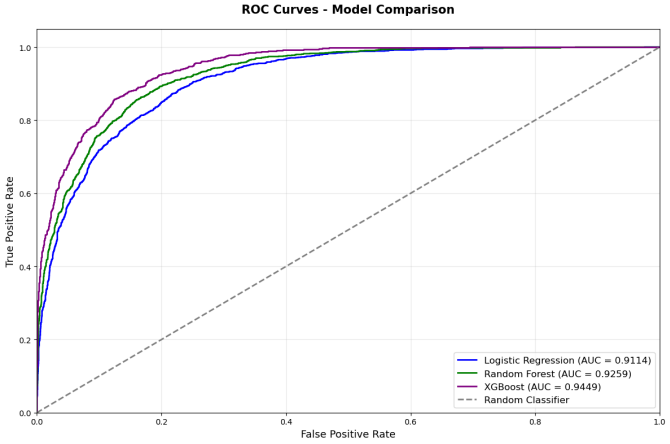

Fig. 7. ROC-AUC metrics after calibration and optimization are shown, comparing the performance of the models chosen

unexpected defeats, which helped in understanding highly correlated features.

These baselines demonstrated that nonlinear relationships contributed marginal gains, motivating further exploration of boosting and imbalance optimization strategies.

C. Feature Importance and Explainability

Feature ranking was derived using correlation analysis and XGBoost's built-in importance metrics. The most predictive variables were linked to **experience, activity, and recent form differences rather than raw rankings**. The top predictors included:

1) *experience_gap_pct:* : career experience difference
2) *activity_diff:* : recent matches played
3) *sectionform_diff_90d:* : Recent 3-month form difference

Negative correlations indicated that as the underdog's relative experience and recent activity increased, the likelihood of an upset rose. Surface-related variables such as *surface_advantage and surface_comfort_diff* also ranked highly, highlighting the contextual nature of tennis performance.

These features can be observed in Fig 8, where interpretability was reinforced, confirming that form and surface mismatches were dominant determinants of upsets.

D. Model Calibration and Class Imbalance Optimization

Since only over 25% of matches were upsets, class imbalance posed a major challenge. We experimented with three strategies:

- **Class Weights:** Penalized misclassification of minority (upset) samples, yielding a 17% recall improvement.
- **SMOTE:** Synthetic Minority Oversampling generated realistic upset cases, improving recall by 15% and overall F1-score by 6.5%.
- **Threshold Tuning:** Adjusting classification probability cutoffs maximized sensitivity (recall up to 98%) at the cost of precision.

Among these, **SMOTE achieved the best balance between recall and precision**, establishing it as the preferred imbalance correction method (Fig 6).

E. Results and Model Comparison

Model evaluation revealed that **XGBoost consistently outperformed other approaches** in discriminative ability and stability. Table I summarizes the key metrics.

TABLE I
MODEL PERFORMANCE COMPARISON ON HELD-OUT TEST DATA.

Model	F1-Score	Recall	Precision	ROC-AUC
LR - Baseline	0.683	0.650	0.720	0.896
RF - Baseline	0.688	0.638	0.747	0.914
XGBoost Baseline	0.745	0.707	**0.787**	**0.945**
XGBoost SMOTE	**0.751**	0.764	0.739	0.938
XGBoost Class Weights	0.749	0.799	0.705	0.935
RF Threshold Tuning	0.590	**0.981**	0.422	0.917

However, best achievements on performance metrics are divided among different models:
– **Best F1-Score**: 0.751 (XGBoost + SMOTE)
– **Best Recall**: 0.981 (Random Forest + Threshold Tuning)
– **Best ROC-AUC**: 0.945 (XGBoost)

These results demonstrate that tennis upsets can be predicted with meaningful accuracy using machine learning and carefully engineered features.

XGBoost-SMOTE confusion matrix is shown (Fig 9) to recall the good performance of these models despite the inherent unpredictability of sports competitions. As we can observe, this model was able to **predict around 76% of upsets, with more than 90% ROC-AUC**.

VI. DISCUSSION & LIMITATIONS

The experimental results reveal several meaningful patterns about the dynamics of tennis upsets and the factors that influence them. The findings not only validate the predictive strength of the models but also offer interpretable insights into the sport itself.

a) Player Experience and Activity.: The analysis confirmed that experience and recent match activity are among the most influential predictors of upsets. Underdogs who have played more matches or maintained consistent competitive rhythm are better prepared to exploit opportunities against higher-ranked opponents.

b) Short-Term Form vs. Long-Term Performance.: Recent performance indicators, particularly 90-day form differentials, proved to be stronger predictors than historical or career-level statistics. This emphasizes that short-term momentum and confidence carry substantial weight in determining match outcomes. From a practical standpoint, this insight aligns with coaching intuition: players entering tournaments with strong recent performances are more capable of challenging top seeds.

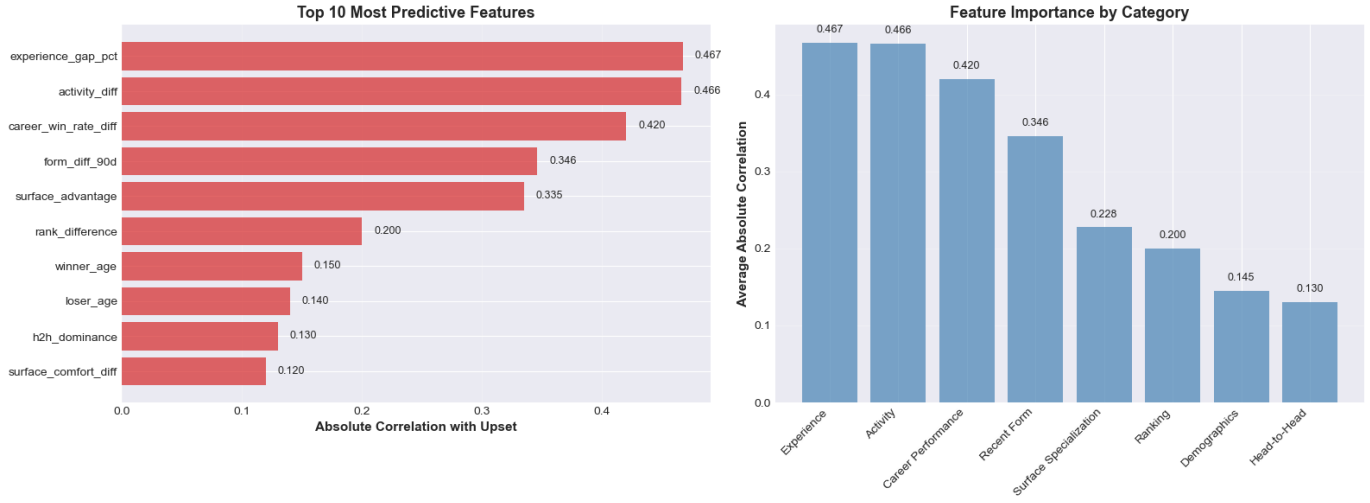

Fig. 8. Most predictive features by absolute correlation with target and feature importance by category

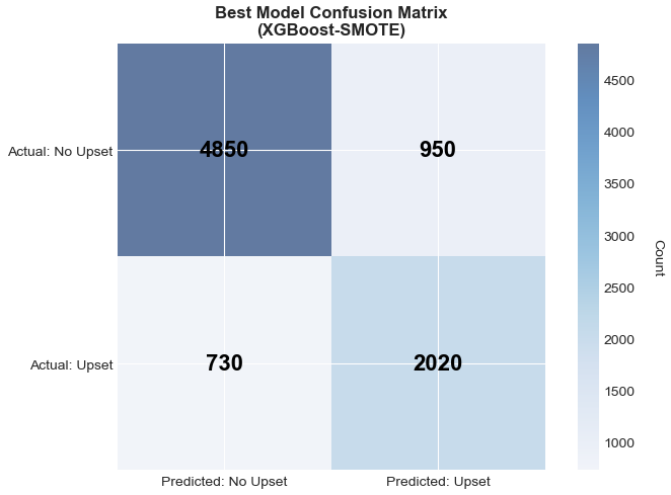

Fig. 9. Best model confusion matrix (XGBoost-SMOTE)

c) Surface and Contextual Factors.: Surface-related variables ranked among the top predictors, underscoring the importance of contextual adaptation. An underdog's familiarity with the playing surface can offset technical disparities, especially when the favorite struggles under specific surface conditions. .

d) Rankings and Head-to-Head Records.: While player rankings and head-to-head history were included in the feature set, their predictive contribution was relatively modest. Rankings fail to account for situational variability, such as fatigue, motivation, or match-specific tactical mismatches. Likewise, head-to-head dominance, although psychologically relevant, showed limited statistical impact—likely due to the large number of first-time encounters in professional tennis.

Overall, the study illustrates that even in a domain dominated by uncertainty, machine learning can extract meaningful and interpretable signals. The discovered patterns mirror intuitive tennis knowledge while quantifying their relative importance, bridging data-driven analytics with real-world decision-making in the sport.

LIMITATIONS:

Besides all the insights found, it is important to talk about several limitations affecting this study:

First, the dataset was limited to ATP-level matches and did not include contextual factors such as player injuries, weather, fatigue, or psychological state, all of which can strongly influence match outcomes.

Second, player performance is inherently dynamic and evolves over time. Even though time-aware validation was applied, non-stationarity in player form or strategic trends can reduce long-term model reliability.

Finally, the tennis domain naturally suffers from class imbalance—upsets are less frequent than expected results. Although techniques like SMOTE and class weighting improved recall, they may introduce minor distortions in the true data distribution.

VII. Conclusion & Future Work

Our experiments demonstrate that machine learning can effectively uncover patterns in the inherently uncertain domain of professional tennis. Despite the sport's stochastic nature, feature-rich boosted models—particularly XGBoost combined with SMOTE and probability calibration—achieved reliable upset probability estimates and interpretable patterns. These results confirm that experience, recent form, and surface specialization are key determinants of unexpected outcomes.

The observed performance ceiling near 75–80% F1-score reflects the natural unpredictability of human performance and the influence of unmeasured psychological and situational factors. Nevertheless, the study highlights that, with appropriate data preprocessing and imbalance handling, meaningful predictive signals can be extracted even from noisy sports data.

Future work should aim to expand both the temporal and contextual depth of the dataset. Integrating real-time match statistics, weather conditions, and player fatigue indicators could capture dynamic performance shifts more accurately. Additionally, incorporating social media sentiment or behavioral metrics may help approximate psychological readiness before matches. From a methodological perspective, exploring temporal deep learning models such as LSTM networks or transformers could better model evolving player form, while causal inference frameworks may help distinguish correlation from genuine influence.

Ultimately, this research provides a foundation for data-driven understanding of tennis upsets and illustrates the broader potential of machine learning in sports analytics—where prediction accuracy and interpretability must coexist within a fundamentally unpredictable environment.

REFERENCES

[1] M. De Seranno, "Predicting ATP Singles Matches Using Machine Learning," 2021. [Online]. Available: https://libstore.ugent.be/fulltxt/RUG01/002/945/727/RUG01-002945727_2021_0001_AC.pdf

[2] J. Dryja, "Machine Learning Approaches for Grand Slam Match Predictions," 2025. [Online]. Available: https://www.cs.vu.nl/~wanf/theses/dryja-bscthesis.pdf

[3] X. Gao and R. Kowalczyk, "Feature-based Prediction of Tennis Match Outcomes Using Random Forests," 2019. [Online]. Available: https://arxiv.org/abs/1910.03203

[4] Y. Bai, et al., "Rhythms of Victory: Predicting Professional Tennis Matches Using Machine Learning," 2024. [Online]. Available: https://www.researchgate.net/publication/383161788

[5] Y. Zhai and X. Wang, "Lasso-Ridge-Based XGBoost for Tennis Outcome Prediction," 2024. [Online]. Available: https://arxiv.org/abs/2405.07030

[6] H. Rui, et al., "Predicting Point Outcomes in Tennis Matches Using Machine Learning," 2024. [Online]. Available: https://proceedings.mlr.press/v245/rui24b.html

[7] A. Author, et al., "AI4Sci Methods for Real-Time Tennis Match Prediction," *Journal of Big Data*, 2025. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-025-01216-4

[8] K. Rosenfield, "Predicting Tennis Outcomes for Betting Using Python," 2021. [Online]. Available: https://kevinrosenfield.com/Predict-Tennis/

[9] J. Sackmann, "ATP Match Results," GitHub repository: https://github.com/JeffSackmann/tennis_atp, 2022

[10] "Searching for the GOAT of tennis win prediction" J. Quant. Anal. Sports, vol. 12, no. 3, pp. 127–138, 2016. Available: https://vuir.vu.edu.au/34652/1/jqas-2015-0059.pdf

[11] J. C. Yue, E. P. Chou, M.-H. Hsieh, and L.-C. Hsiao, "A study of forecasting tennis matches via the Glicko model" PLoS ONE, vol. 17, no. 4, art. no. e0266838, Apr. 2022. Available: https://journals.plos.org/plosone/articleid=10.1371/journal.pone.0266838

[12] J. Lou et al., "Comparative Analysis of Logistic Regression, Random Forest, and XGBoost", Atlantis Press, 2024. Available: https://www.atlantis-press.com/article/126004036.pdf

[13] S. Fatima, A. Hussain, S. B. Amir et al., "XGBoost and Random Forest Algorithms: An In Depth Analysis". Available: https://www.researchgate.net/publication/377135877_XGBoost_and_Random_Forest_Algorithms_An_in_Depth_Analysis

[14] M. Imani, A. Beikmohammadi, H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance ...", Technologies, 2025. Available: https://www.mdpi.com/2227-7080/13/3/88

[15] Tennis Match Prediction Based on Wavelet-BP Neural Network Model. Available: https://www.researchgate.net/publication/382567409_Tennis_Match_Prediction_Based_on_Wavelet-BP_Neural_Network_Model

[16] GitHub Repository for the project, including all the coding notebooks and visualizations. Available: https://github.com/Manurguez03/Comp.-Data-Analysis-Repository