ANALYZE TENNIS WINNING FACTORS ACROSS DIFFERENT SURFACES BY UTILIZING RANDOM FOREST

WUHUAN DENG

Department of Applied Mathematics, University of Washington, Seattle, WA
briandeng0216@gmail.com

Abstract. Tennis is one of the most popular sports worldwide, with a rich calendar of professional tournaments played across three court surfaces: hard, grass, and clay. Each surface has unique physical characteristics that significantly influence ball behavior, player movement, and match dynamics. As a result, different playing styles tend to be more effective on certain surfaces. This research investigates the surface-dependent nature of match outcomes by exploring statistical trends and performance indicators that contribute to success on each court type. Understanding these differences can provide deeper insights into player adaptability, match strategies, and surface-specific training.

1. Introduction

1.1. **Motivation.** Grass, clay, and hard courts differ significantly from one another, each possessing unique characteristics that influence gameplay. Grass courts are the most traditional surface, featuring low, unpredictable bounces and a fast pace. They are preferred by players who favor a serve-and-volley style [Nag, 2022a]. Roger Federer is widely regarded as the greatest player on grass, having won eight Wimbledon men's singles titles between 2003 and 2017 [**?**].

Clay courts are made of crushed stone and other minerals. They produce higher bounces and slow down the ball, making it more challenging to hit winners [Nag, 2022a]. This surface tends to benefit players who excel in long rallies. Rafael Nadal, also known as the "King of Clay," has won the most men's singles titles at Roland Garros, with 14 championships [Chennai, 2024].

Hard courts are the most common surface in modern tennis. The speed of play on hard courts varies depending on the composition of the top layer, but in general, they are faster than clay courts and slower than grass courts [Nag, 2022a]. The US Open and Australian Open—two of the four Grand Slams—are played on hard courts. This surface is often favored by all-around players without significant weaknesses, such as Novak Djokovic [Nag, 2022a].

Given this information, different techniques and strategies are required to succeed on each surface. In this research, we are going to analyze professional match data from grass, clay, and hard courts to examine how various match features influence winning probabilities on each surface.

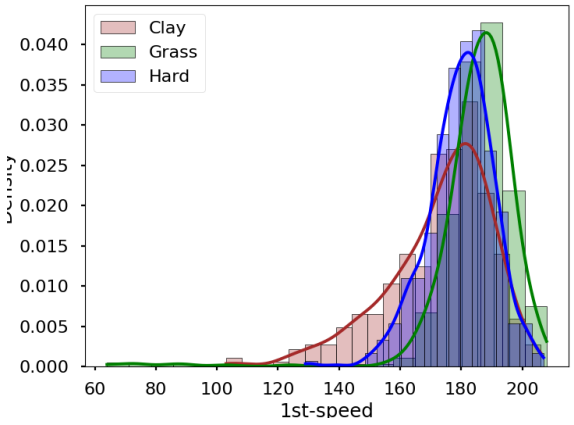

FIGURE 1. Distribution of 1st-serve speed across the three surfaces. Grass courts generally yield the fastest serve speeds, while clay courts have the lowest.

1.2. **Related Works.** Soomedha Vasudevan and Nick Chu analyzed how various match features influence winning percentage across the three surfaces using linear regression methods [Vasudevan and Chu, 2023]. They computed the R-squared values between individual features and surface-specific win percentages, finding that the same feature could have slightly different R-squared values depending on the surface. However, most of these R-squared values were close to zero, and the differences across surfaces were minimal. Moreover, their analysis did not reveal how changes in feature values influence win probability, beyond simple scatter plots. Our research extends this work by using Random Forest classifiers with multiple input features, enabling nonlinear modeling and capturing interactions between variables. By incorporating feature importance and partial dependence analysis, our method uncovers more nuanced and surface-specific patterns in match outcomes that linear models may overlook.

2. MATERIALS AND METHODS

2.1. **Dataset.** The match data used in this research was collected from GitHub [Sackmann, 2024]. The dataset includes matches from three Grand Slam tournaments: the US Open (2018, 2019, 2023, and 2024), Wimbledon (2021, 2022, 2023, and 2024), and the French Open (2015 and 2016). In total, the dataset contains 479 matches played on grass, 437 on hard courts, and 187 on clay

courts. The data is structured on a point-by-point basis, with each row corresponding to a single point and each column representing a specific match feature.

2.2. **Random Forest.** Random Forest is a widely used supervised learning algorithm for both classification and regression tasks. It is composed of multiple decision trees and outputs a class label for classification problems or an average prediction for regression problems [Breiman, 2001]. Each tree in the forest is trained on a bootstrap sample of the data, and at each node split, a random subset of features is considered to promote diversity among trees. Given an input x and a Random Forest consisting of n trees, the predicted output is defined as:

$$\hat{y} = \frac{1}{n}\Sigma_{t=1}^{n} f_i(x),$$

where $f_i(x)$ is the prediction from the ith decision tree. By ensembling multiple trees, Random Forest reduces variance and helps prevent overfitting compared to using a single decision tree.

In this research, three separate Random Forest models are trained—one for each court surface. The input to each model consists of 10 player-specific features, and the output is a binary classification indicating the match outcome: 0 for a loss and 1 for a win.

Feature Name	Description
1st-speed	Average first serve speed
2nd-speed	Average second serve speed
ace-rate	Number of ace / Number of serves
double-fault-rate	Number of double faults / Number of service points
1st-rate	Number of 1st serve in / Total 1st serve attempts
1st-win-rate	Points won on 1st serve / Number of 1st serve in
2nd-win-rate	Points won on 2nd serve / Number of 2nd serve in
net-win-rate	Points won at net / Number of net approaches
winner-rate	Number of winners / Total points played
unforced-error-rate	Number of unforced errors / Total points played

TABLE 1. This table lists the 10 input features and their basic descriptions. For clarification: an ace is a serve that lands in the service box and is not touched by the returner; a winner is any shot that lands in bounds and is not returned by the opponent; and an unforced error is a mistake made when the player is not under pressure—in other words, an error that should be avoidable at the professional level.

Surface	n estimators	max depth	min samples leaf	bootstrap
Grass	100	10	5	True
Clay	50	5	5	True
Hard	100	10	5	True

TABLE 2. This table shows the Random Forest parameter values used for each surface. All other parameters not listed here are set to their default values from the *scikit-learn* package.

2.3. **Feature Importance.** To assess the importance of each feature in predicting match outcomes using the Random Forest model, we computed feature importance based on Mean Decrease in Impurity (MDI) [Breiman, 2001]. This metric quantifies the total reduction in node impurity attributed to each feature across all trees in the forest. Given a feature x_j, its importance $I(x_j)$ is defined as:

$$I(x_j) = \frac{1}{n}\Sigma_{t=1}^{n}\Sigma_{k\in N_j^{(t)}} p(k)\cdot\Delta i(k),$$

where n is the number of trees, $N_j^{(t)}$ is the set of nodes in tree t where feature x_j is used for splitting, $p(k)$ is the proportion of training samples reaching node k, and $\Delta i(k)$ is the impurity reduction achieved at that node [Louppe et al., 2013]. This quantitative metric provides a ranking of features based on their influence on the model's final predictions. The MDI for each feature ranges from 0 to 1, and the sum of all feature importances equals 1.

MDI Value	Importance Level
> 0.15	Very important - dominant feature
0.05 - 0.15	Moderately important - strong signal
0.01 - 0.05	Weak but maybe useful
< 0.01	Not useful

TABLE 3. This table explains the interpretation of each MDI value range in terms of feature importance.

2.4. **Partial Dependence.** To further analyze the impact of each feature, we utilized Partial Dependence (PD) and visualized it across different features and court surfaces. A Partial Dependence Function (PDF) measures the average model prediction as a function of one or more selected input features, while averaging out the effects of all other features [Friedman, 2001]. Given a prediction function $f(x)$ trained by a Random Forest model, and a subset of features S, the PDF is defined as:

$$\hat{f}_s(x_s) = \frac{1}{n}\Sigma_{i=1}^{n} f(x_s, x_C^{(i)}),$$

where C is the complement of S, and $x_C^{(i)}$ represents the values of the remaining features for the ith instance in the dataset. This metric estimates the expected model output when the features in S are fixed, while the other features vary according to their observed distribution. For example, if a player has a 1st-win-rate of 0.75 and the corresponding partial dependence value is 0.76, it means that if all players had a 1st-win-rate of 0.75, the model would predict an average win probability of 76%, regardless of the other feature values.

3. Result

3.1. **Random Forest.** To ensure that the computed feature importance and partial dependence values are meaningful and reliable, it is essential to begin with well-performing Random Forest models. Therefore, we require all three models—one for each surface—to achieve at least 80% accuracy on the testing set. The testing data consists of 20% of the total matches for each surface, randomly selected from the dataset.

Surface	Training Accuracy Score	Testing Accuracy Score
Grass	0.925%	0.904%
Clay	0.934%	0.819%
Hard	0.906%	0.833%

Table 4. This table shows Random Forest model performance results for each court surface.

3.2. **Feature Importance.** Next, we computed the feature importance scores for each input feature using the trained Random Forest models.

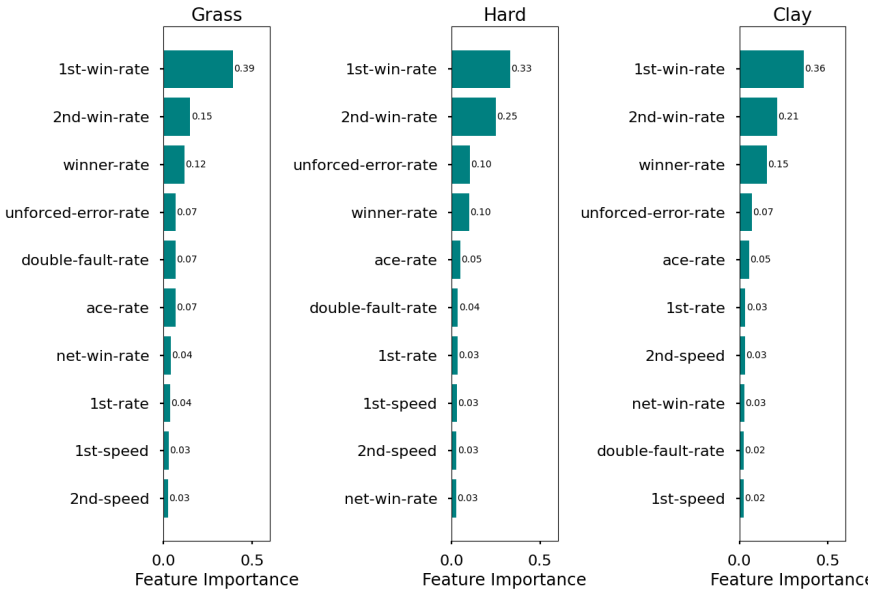

FIGURE 2. Feature importance of each model.

Unsurprisingly, 1st-win-rate is the most influential feature across all three surfaces, with importance scores of 0.36 on grass, 0.33 on hard, and 0.36 on clay. This aligns with both our expectations and traditional tennis insights from players and coaches—capitalizing on first-serve opportunities is widely considered one of the most critical factors for match success. The 2nd-win-rate ranks second in importance, although its contribution on grass courts is notably lower compared to hard and clay surfaces. This observation is reasonable, as every point begins with a serve—whether first or second—and the ability to consistently win serve points strongly correlates with overall match outcomes. Given that 1st-win-rate and 2nd-win-rate dominate the model, it becomes especially important to examine the remaining features to uncover more nuanced, surface-specific patterns.

The third most important feature on hard courts differs from that on grass and clay courts. Winner-rate ranks third in importance on both grass and clay surfaces, whereas unforced-error-rate holds the third position on hard courts. This may suggest that aggressive playing styles—which often result in more winners—are more advantageous on grass and clay courts. This is particularly true for clay courts, where the slower surface makes it harder to hit winners; thus, players who can still generate them may gain a significant competitive edge.

Although winner-rate is not ranked third on hard courts, its importance score (0.10) is equal to that of unforced-error-rate, suggesting a balanced contribution from both features. This supports the earlier observation that all-around players tend to perform better on hard courts, where no

single playing style is dominant. In comparison, the importance of unforced-error-rate on grass and clay courts is lower, at 0.07 for both. This implies that while unforced errors are still relevant, taking calculated risks that may result in winners can be more rewarding on grass and clay surfaces.

Focusing on serve-related features, grass courts show the highest feature importance for both ace-rate and double-fault-rate, each with a value of 0.07. This highlights the critical role of serving on grass—players are rewarded not only for their ability to generate aces but also for their ability to minimize errors. In contrast, double-fault-rate has a much lower importance on clay courts, with a value of just 0.02, suggesting that serve-related mistakes are less consequential on slower surfaces like clay.

Ace-rate holds moderate importance on both clay and hard courts, with a value of 0.05 for each. While serving may not offer as significant an advantage as it does on grass courts, it remains a critical component of success across all surfaces. These results indicate that although grass courts demand the most from serve performance, the ability to serve effectively continues to play an important role in winning matches on any surface.

3.3. **Partial Dependence.** To further explore how each feature influences winning probability, we conducted a detailed analysis of several selected features.

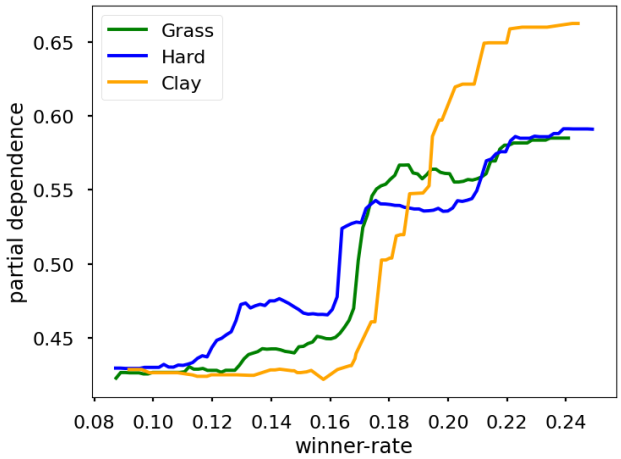

FIGURE 3. Partial dependence of winner-rate on winning probability across the three court surfaces.

From the graph above, it is evident that a higher winner-rate consistently correlates with increased winning probabilities across all three court surfaces. When the winner-rate is below 0.12, the predicted winning probability remains low and relatively flat. On hard courts, the winning

probability begins to rise once the winner-rate exceeds 0.13, indicating that even a slight increase in aggressive play can lead to better outcomes. Grass courts follow a similar pattern, with a noticeable increase occurring at a comparable threshold. In contrast, the clay court curve rises more gradually at first but shows a sharp increase once the winner-rate surpasses 0.16.

When the winner-rate reaches 0.20, the winning probabilities on both hard and grass courts begin to plateau. However, on clay courts, the upward trend continues sharply until approximately 0.22. Beyond this point, the growth levels off, but clay courts maintain the highest predicted winning probability when the winner-rate exceeds 0.20. This pattern suggests that on slower surfaces like clay, the ability to generate winners has a particularly significant impact on match outcomes.

Overall, increasing the winner-rate from 0.12 to 0.24 leads to at least a 0.15 increase in predicted winning probability across all surfaces. This underscores the critical role of hitting winners in achieving match success, regardless of the court surface.

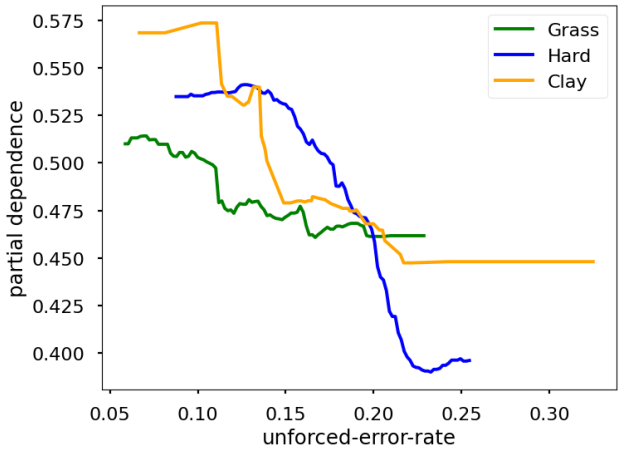

FIGURE 4. Partial dependence of unforced-error-rate on winning probability across the three court surfaces.

From the graph above, we observe that achieving a winning probability above 0.5 requires a very low unforced-error rate—below 0.1—across all three surfaces. As the unforced-error-rate increases, the winning probability consistently decreases, with the steepest decline occurring on hard courts. In contrast, grass and clay courts exhibit more gradual declines. When the unforced-error rate exceeds 0.2, the winning probabilities on all surfaces fall below 0.45, with hard courts showing the lowest values. This suggests that consistency plays a particularly important role on hard courts, where minimizing unforced errors yields the greatest impact on match outcomes.

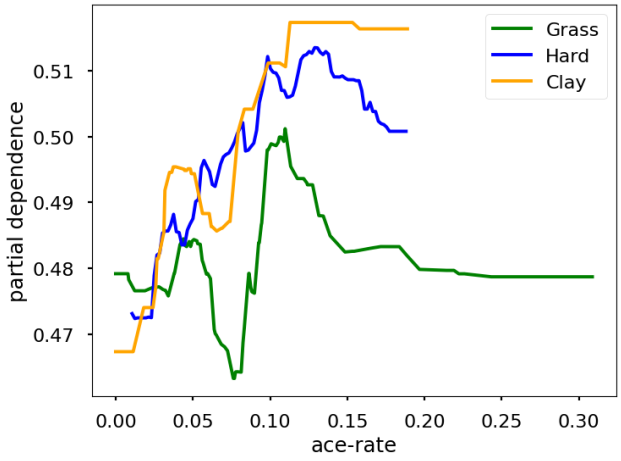

FIGURE 5. Partial dependence of ace-rate on winning probability across the three court surfaces.

From the graph above, the relationship between ace rate and winning probability shows greater variability compared to winner-rate. When the ace-rate exceeds 0.05, both clay and hard courts show an increase in winning probability, while the grass court curve unexpectedly dips before beginning to rise again around 0.075. Between an ace-rate of 0.13 and 0.15, all three surfaces reach their peak predicted winning probabilities, with clay courts showing the highest value. However, beyond 0.15, winning probabilities drop sharply on both hard and grass courts, while the value on clay remains relatively stable. A possible explanation is that aces are more difficult to achieve on clay, the slowest surface; thus, players who manage to hit many aces on clay may be more well-rounded. In contrast, on faster surfaces, players who rely heavily on aces might lack other essential skills—such as groundstrokes—leading to lower overall performance. As previously noted, all-around players tend to have an advantage on hard courts in particular.

4. DISCUSSION

This research currently focuses on individual features. However, in actual matches, many metrics are closely related—for example, aces and double faults, or winners and unforced errors. It would be reasonable to create more informative features by combining related variables and incorporating them into the model. Additionally, although we separated the data by surface type, there are still notable differences between tournaments held on the same surface. For instance, the Australian Open and US Open—both played on hard courts—have different surface characteristics. Therefore, analyzing surface conditions at the tournament level could provide more practical insights for

players preparing to compete. Lastly, our analysis has focused on outcome-based statistics, such as aces, rather than the mechanics of the actions themselves, like serve speed or spin rate. To further improve player performance, it may be beneficial to analyze technical attributes of individual shots.

5. Conclusion

Although it is well known that increasing 1st-win-rate and 2nd-win-rate benefits players, our model results offer additional, previously unseen insights. Hard courts demand well-rounded skills, requiring players to excel in both serving and groundstrokes. On clay courts, where the surface slows down play, players who can still produce aces and winners gain a distinct advantage. For grass courts, effective serving is crucial, but controlling double faults is equally important. We believe our analysis can offer practical suggestions for professional players in preparing their strategies for different surfaces. Of course, these insights alone cannot directly enhance performance—consistent training, both on and off the court, remains essential.

Acknowledgments

I would like to thank the authors of the tennis-slam-pointbypoint GitHub repository for making their dataset publicly available, which provided a valuable foundation for this research.

References

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Chennai, 2024] Chennai (2024). Rafael nadal at french open: Full list of titles won, nadal's record at roland garros before 2024 match against zverev.

[Friedman, 2001] Friedman, H. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

[Louppe et al., 2013] Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). Understanding variable importances in forests of randomized trees. 26.

[Nag, 2022a] Nag, U. (2022a). Everything you need to know about tennis courts.

[Nag, 2022b] Nag, U. (2022b). Roger federer at wimbledon: When 'king roger' ruled.

[Sackmann, 2024] Sackmann, J. (2024). tennis slam pointbypoint.

[Vasudevan and Chu, 2023] Vasudevan, S. and Chu, N. (2023). Decoding surface deominance: The skills behind tennis triumphs.