A Deeper Look into the Gameplay of Tennis: How Psychological Nuances Affect a Player's Performance and a Match's Outcome

Minh Hoang, Zaynab Khan, Xiang Liu, Damien Snyder
University of Washington, Seattle, WA 98105
{minh257,zaynabk,lukaro,damiensn}@cs.washington.edu

Abstract

A professional tennis player's success depends not only on physical skill but also on mental preparation. Spectators often claim that a player tends to choke by performing poorly under pressure, or that a match is streaky because the players had pronounced streaks of good or bad performance. These phenomena are hard to verify because we cannot know players' thoughts from the outside. In this paper, we investigate whether these mental factors can be identified from play-by-play data from Grand Slam tennis matches, and how gameplay changes overall throughout a match. We introduce two measurements of players' tendencies to choke and a method of identifying streakiness in tennis matches, and we quantify changes in gameplay patterns between sets of a match. Our analysis can inform tennis players on the mental side of tennis, and the tennis audience on how and how not to expect mental factors to influence a tennis match.

1 Introduction

Researchers have proposed multiple statistical models [1, 16] in order to predict professional tennis match outcomes accurately. De Serrano et al. [4] have shown that machine learning and statistics techniques may even allow one to make a significant profit on betting markets. However, these methods of predicting match outcomes are based solely on player statistics, which count every point the same, regardless of whether the point occurred in a critical moment or well after the match was out of hand. The ups and downs of a tennis match can be mentally taxing for a tennis player, whose mental state will likely depend in large part on the state of the match. Several well-known tennis players have spoken about mental-related performance issues, including Naomi Osaka [2], Robin Soderling [12] and Nick Kyrgios [5].

Because of the potential applications in match prediction and in tennis' players own psychological preparation, we believe it is useful to study player mentality and how it relates to the state of a tennis match. We will conduct our analyses to answer the following research questions:

- **RQ1:** Are tennis matches streaky?
- **RQ2:** Do some tennis players choke more than others in high-pressure situations?
- **RQ3:** How does tennis gameplay (e.g. the rate of unforced errors, winners, etc.) of players change over the course of a match?

Streakiness refers to the tendency of a player to win points more often than usual if they have performed well recently, and lose points more often if they have performed poorly recently. The goal of **RQ1** is to understand whether streakiness actually exists in tennis, or if it's just a popular misconception, whereas **RQ2** investigates the definition of *choking* in tennis, which refers to the tendency of a player win points less often than usual in important, high-pressure situations. We want to understand what scenarios players tend to choke more and whether these situations apply to all players. Finally, **RQ3** helps us understand gameplay changes over the course of a match, which reveal how a player's mentality is affected as they get into later sets of the match and potential strategies to take advantage of these changes.

2 Dataset

All proposed research questions will be analyzed and evaluated using a total of 3 different datasets that are publicly available [9, 10, 11].

The ATP Tennis Stats dataset [9] contains player names, rankings, tournaments and match results, along with detailed match statistics of every ATP tennis match from 1968-2022. This dataset acts as a baseline and some of its features will be combined with the other 2 datasets during preprocessing to gain more depth for our analysis of each research question.

The Grand Slam Point-by-Point dataset [10] includes over 1 million point-level data points of every Grand Slam match between 2011 and 2022. This dataset will be used to measure the chokiness of players in Grand Slams during this period.

The Match Charting dataset [11] contains point-by-point data of several Men's Singles matches spanning from 1968-2022 in greater details compared to the Grand Slam Point-by-Point dataset, including the type of shot, direction of shot, depth of returns, types of errors, and aggregate match-level data. This dataset is used to analyze our questions about streakiness and gameplay changes.

Dataset	**Rows**	**Max. Features**	**Categorical Features**	**Matches**	**Players**
Grand Slam Point-by-Point	1,020,105	85	23	5,461	448
Match Charting	619,647	30	13	3,599	660

Table 1: *Distribution of data points in the Grand Slam Point-by-Point and Match Charting datasets. Data only includes Men's Singles matches.*

Due to the large size and differing time span of each dataset, we focus on one specific span where the data is most complete: 2011-2022. Several works have been conducted using a smaller range of data, usually between 3 to 5 years, but our method uses data from an 11 year span between the beginning of 2011 to the end of 2021, focusing on Men's Singles matches, in order to provide a wider range of analysis.

Before conducting our analyses, some data manipulation was necessary. First, match information outside of the time range and demographic (Men's Singles) are removed. Then, format inconsistencies in columns such as players' names and match IDs are converted into a consistent format across all tournaments and years. Finally, we combine the data into larger tables based on the year of the tournament and drop all features that are more than half-empty (containing NAN values). For features that we consider essential for our analysis but are not complete, we put placeholder values to avoid imbalance between years.

3 Analysis Approach

We propose the following hypotheses and use our data to validate them:

- **Hypothesis 1 (H1):** Streakiness exists in tennis.
- **Hypothesis 2 (H2):** Some players, notably more successful players (e.g. Grand Slam winners) choke less than other players.
- **Hypothesis 3 (H3):** Over the course of a match, the likelihood of a player hitting an error will increase, and the likelihood of hitting a winner will decrease.

3.1 Hypothesis 1: Streakiness exists in tennis

Validating **H1** will help us answer **RQ1**. We use the Match Charting Project data to determine whether a player will win the next point based on their previous points. Specifically, for each point

P in our dataset, we focus on the *server* - the player who serves in that point. We keep track of the number of points they have played and won up until point P, both when serving and receiving. We also consider the following binary variables:

S_i: Whether the server won the i-th most recent point they served, for $i \in [1, 5]$
R_i: Whether the server won the i-th most recent point they received, for $i \in [1, 5]$

From these variables, we compute two additional features of the server for each match:

$$W_S = \frac{\sum_{i=0}^{P-1} \text{(Points won when serving)} + a}{\sum_{i=0}^{P-1} \text{(Points played when serving)} + c} \qquad \text{(Win percentage when serving)}$$

$$W_R = \frac{\sum_{i=0}^{P-1} \text{(Points won when receiving)} + c - a}{\sum_{i=0}^{P-1} \text{(Points played when receiving)} + c} \qquad \text{(Win percentage when receiving)}$$

These two features are the server's win percentages up to point P in the match when they serve and receive. In our overall analysis of the data, the server won 64% of the total points. Therefore, we use Laplacian smoothing to make the win percentages less noisy, using $c = 50$ smoothing points and assume the server wins $a = 32$ of them.

Let p be an indicator variable for "the server wins point P". We quantify streakiness by comparing two predictive models for p. The first model, M_1, only has 2 inputs: W_S and W_R, while the second model, M_2, has 12 inputs: W_S, W_R, S_i, and R_i for $i \in [1, 5]$. We exclude all points where the server has not previously served at least 5 points. We use these models to predict the outcome of each point in our data and compare the mean-squared error of our two models. Our data is split into training and testing with a ratio of 75:25. For our models, we use Linear and Logistic Regression. Each model is trained on 2 cases: With and without knowing the order of the last 5 points.

3.2 Hypothesis 2: Choking between different players

We will analyze **H2** to answer **RQ2** using the Grand Slam Point-by-Point and ATP Tennis Stats datasets to determine if some players tend to choke more than others during high pressure situations. We define *high pressure situations* as when a player is facing a *break point* - the state of the game when the receiver needs only one more point to win the server's game. Other points will be considered *normal points*. To validate our hypothesis, we will calculate the *choking percentage* of a player. We define *choking* as when the player hits an unforced error during high pressure moments. For simplicity, we consider *double faults* - when a player makes both first and second serve faults in a point, thus losing that point - to be unforced errors as well.

In order to calculate the choking percentage, for each point P of a match, we need to keep track of the following features: The number of normal points the server and the receiver have played and won up to point P, the number of break points the server has played and saved up to point P, and the number of break points the receiver has played and won up to point P. We define and calculate the server and receiver's percentage of winning normal points and break points up to point P as follows:

$$S_N = \frac{\sum_{i=0}^{P-1} \text{(Normal points server won)} + a}{\sum_{i=0}^{P-1} \text{(Normal points played)} + c} \qquad \text{(\% of server winning normal points)}$$

$$R_N = \frac{\sum_{i=0}^{P-1} \text{(Normal points receiver won)} + (c - a)}{\sum_{i=0}^{P-1} \text{(Normal points played)} + c} \qquad \text{(\% of receiver winning normal points)}$$

$$S_B = \frac{\sum_{i=0}^{P-1} \text{(Break points server saved)} + b}{\sum_{i=0}^{P-1} \text{(Break points played)} + c} \qquad \text{(\% of server saving break points)}$$

$$R_B = \frac{\sum_{i=0}^{P-1} \text{(Break points receiver won)} + (c - b)}{\sum_{i=0}^{P-1} \text{(Break points played)} + c} \qquad \text{(\% of receiver winning break points)}$$

We also use Laplacian smoothing to reduce the noise of the percentages, where a is the smoothing coefficient of winning as a server over c smoothing points, and b is the smoothing coefficient of winning break points over c smoothing points. In our case, we choose $a = 64, b = 60, c = 100$ because the server wins about 64% of normal points and 60% of break points in our data.

We calculate the choking percentage of a player when serving and receiving up to point P as follows:

$$S_C = S_N - S_B - 0.02 \qquad \text{(Choking percentage as server)}$$

$$R_C = R_N - R_B + 0.04 \qquad \text{(Choking percentage as receiver)}$$

We include the constants 0.02 and 0.04 because servers win break points at a lower rate than non-break points. This difference skews the choking percentages, so the constants adjust all choking percentages such that the average is closer to 0. The *choking percentage* of a player in a match is then defined as the average of their serving and receiving choking percentages.

In addition to our approach, we also apply Harris et al. [6]'s approach for analyzing choking to our dataset, which defines a point to be a high pressure situation if it belongs to these scenarios:

- If the point is late in the game, where the score is at least 30 - 30.
- If the opponent is one game away from winning a set (For example: 5 - 1, 5 - 4, or 6 - 5).
- If the point is played in the deciding set (For Men's Singles in Grand Slams, the fifth set).
- If the opponent is having *game points* - the state of the game when the opponent needs only one more point to win the game - either as the server or receiver.
- If facing break points.

We define a new feature called *Pressure Points*, where each pressure point is incremented for a point P if it belongs to one of the situations above. A point can belong to none or multiple high pressure moments, so pressure points can have a minimum of 0 and maximum of 5. An example of points with a pressure of 0 are the first 2 points of every match, and a point with a pressure of 5 is when a player is down 40 - 30 on their serve (facing a break/game point - in this case, match point - late in the game), in the deciding set, and is currently down by 6 games to 5 (the opponent is one game from winning the set - in this case - the match).

We also define and keep track of another feature called *Post Error Points*, an indicator variable for "point P is preceded by an unforced error in the same set":

$$PostError(P_x) = \mathbb{1}(P_x) := \begin{cases} 1 & \text{if } Error(P_{x-1}) = 1 \text{ and } SetNo(P_x) = SetNo(P_{x-1}) \\ 0 & \text{otherwise} \end{cases}$$

Adapting from Harris et al., we use a Generalized Linear Mixed - Effects Model L_1 to measure the choking percentage of a player during different levels of high pressure situations. A Generalized Linear Mixed - Effects Model (GLMM) [15] is an extension of the Generalized Linear Model (GLM) - a more flexible generalization of the ordinary Linear Regression where the linear predictor contains random effects in addition to fixed effects. Our model, L_1, is as follows:

$$L_1 = \textit{ChokingPercentage} \sim \textit{PressurePoint * PostError + (1 | PlayerID)}$$

In the formula, *ChokingPercentage* is the response, *PressurePoint, PostError* are the independent variables and *PlayerID* is the random effect. L_1 is run to examine the effects of multiple high pressure situations, combined with post error points, on a player's choking percentage. To validate our results, we will analyze the *Odd Ratios (OR), Confidence Interval (CI)* and *p-values* of our model.

3.3 Hypothesis 3: The change in probability of unforced errors and winners over time

We study **H3** to answer **RQ3** using the Grand Slam point-by-point and Match Charting datasets. For each set i of the matches in our data, we keep track of the following variables: The number of unforced errors and winners a player has won in set i, as well as the total number of points played in that set. Similar to our approach in **H2**, we consider *double faults* to be unforced errors, and *aces* - serves that successfully lands in the service box and does not touch the receiving player's racquet - to be winners as well.

From these variables, we can calculate the following for each player:

$$E_i = \frac{\text{Unforced errors in } SetNo(i)}{\text{Points played in } SetNo(i)}, \text{ for } i \in [1, 5] \qquad \text{(Percentage of unforced errors of each set)}$$

$$W_i = \frac{\text{Winners in } SetNo(i)}{\text{Points played in } SetNo(i)}, \text{ for } i \in [1, 5] \qquad \text{(Percentage of winners of each set)}$$

We also use GLMMs to model the effects of set number and post error points on the probability of errors, winners, and net points. We define the following 2 GLMMs:

$$L_2 = E \sim \mathit{SetNo} + (1 \mid \mathit{PlayerID})$$
$$L_3 = W \sim \mathit{SetNo} + (1 \mid \mathit{PlayerID})$$

Each model will also be validated using *OR, CI* and *p-values*.

4 Results and Findings

4.1 Hypothesis 1: Streakiness exists in tennis

We find that knowing the winners of the most recent points in a match only improved our models' predictive accuracy by a minuscule amount. This holds for both models with differing amounts of Laplacian smoothing and recent point inclusion.

We score models by having each of them predict whether the server won a point and computing loss using mean squared error (which is identical to Brier score in our case). So if a model outputs a 70% chance that the server wins and the server loses, the loss on that point is $(0 - 0.7)^2 = 0.49$. If the server wins, the loss would instead be $(1 - 0.7)^2 = 0.09$. Table 2 shows the mean squared error loss on our test dataset for each model.

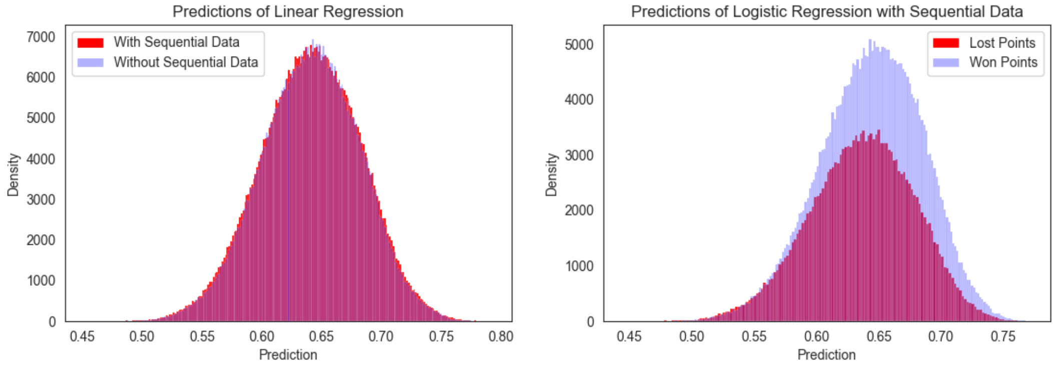
Inputs included	**Linear**	**Logistic**
Recent points and win percentage (M_2)	0.22727	0.22726
Only win percentage (M_1)	0.22729	0.22729
Only recent points	0.22831	0.22830

Table 2: *Mean squared error losses of models M_1 and M_2 with different inputs.*

As a point of comparison, predicting the same value (about 64%) regardless of input gives a loss of at minimum 0.22891. While the predictive models are an improvement on naïvely predicting the same value every time, and M_2 is an improvement over M_1, the improvements are very small, suggesting that the influence of streakiness in tennis is very slight.

For the linear regression models, Figure 1a shows that the predictions are never extremely confident in one direction or another. Most center around the mean value, 64%, and nearly all predictions differ from that value by 15% or less. Our other models had comparable distributions of predictions.

For the logistic regression models, Figure 1b demonstrates how slight the usefulness of the predictive models are, compare the distribution of its predictions on points the server won (blue) versus its predictions on points the server lost (red).



(a) *Probability densities of Linear Regression models* **(b)** *Probability densities of Logistic Regression models*

Figure 1: *Probability densities of different regression models. Figure 1a shows that there's not much difference in predictions of linear regression models that does and does not know the sequential data. Figure 1b shows a slight difference in the distribution of predictions of logistic regression models on points the server won and lost.*

The distributions are nearly identical, with won points having predictions only slightly more in favor of the winner. This pattern held both when we reduced the number of recent points included from 5 to 3 and increased the number of recent points included from 5 to 10. Increasing the recent points history from 5 to 10 did not substantially change the difference in accuracy between M_1 and M_2. Likewise we found that amounts of *ghost points* used in Laplacian smoothing other than 50 reduced predictive accuracy slightly compared to 50.

Finally, we inspect the coefficients of the models to verify whether they match our intuitions for what would positively predict success on a point. We find that this generally is the case: the largest coefficient in each model is the server's win percentage, W_S, when serving, followed by their win percentage, W_R, when receiving. This is followed by the outcomes of the most recent points, S_i, the server served on, and finally the outcomes of the most recent points, R_i, the server received on. This aligns with our intuitions because win percentage is more holistic than the outcome of a single point and because a player's success when serving is probably more correlated with his previous success when serving than with his previous success when receiving.

Based on our results, we can see that there are no substantial differences in predicting the next point's winner whether we know the sequences of recent points or not. Therefore, for our current approach, we reject **H1** and conclude that streakiness does not exist in tennis.

4.2 Hypothesis 2: Choking between different players

We filter our dataset to visualize players that have faced at least 100 break points in all matches. Among the filtered players, we separate them into 2 clusters: Grand Slam winners and others. Figure 2 shows the distribution of choking percentages of players, with all Grand Slam champions during the 2011 - 2021 period highlighted. From our observation, the choking percentages of the majority of Grand Slam champions, including more than 10-time-winner Novak Djokovic, are higher than average, which we find hard to believe given top players' reputations for mental toughness. We investigate these results to see if they might be due to random chance, and we find that the variance in the distribution of choking percentages is approximately equal to what would be expected if no players had any tendency to choke (i.e., the effects of only random noise).

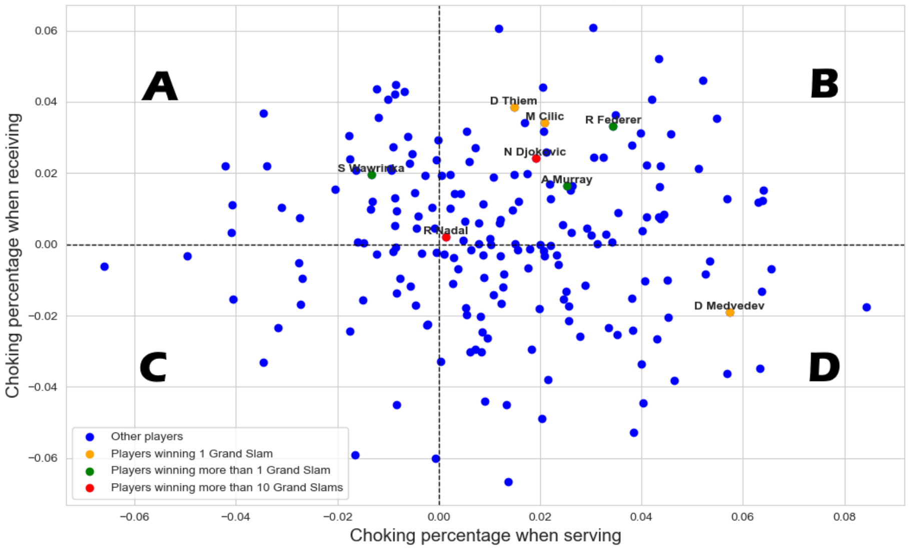
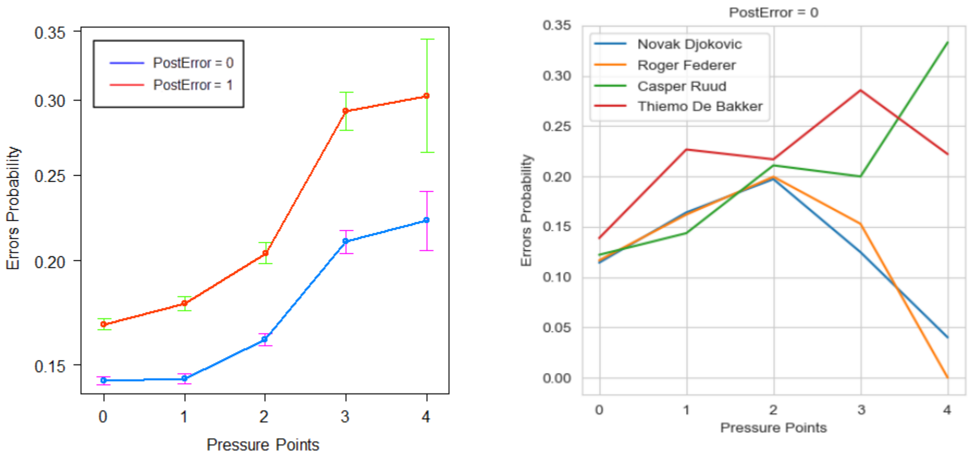


Figure 2: *Choking percentage of players when facing break points. Highlighted players are Grand Slam winners of different quantities during 2011 - 2021. Section A includes players who choke more when receiving but less when serving. Section D includes those who choke more when serving but less when receiving. Section B consists of players tending to choke a lot both when serving and receiving, and section C are players who are less likely to choke in both situations.*

In our second approach, we only consider choking percentage as the likelihood of hitting an unforced error during high pressure situations. Additionally, the number of points having pressure points equal to 5 is very low, so we group them with points when pressure points is 4.

Figure 3a shows the results of our L_1 model: the mean probability of players hitting an unforced error during different levels high pressure situations. According to Figure 3a, the likelihood of hitting an error is positively correlated with pressure. With the inclusion of Post Error points, the overall likelihood of hitting an error increases significantly. These claims are further strengthened by the statistical summary of our model L_1 in Table 3, where the Odds Ratios (ORs) of predictors are greater than 1, and their associated Confidence Intervals (CIs) are relatively small, which means that the probability of hitting an error is likelier to occur as the value of our predictor increases. Moreover, since their p-values are mostly less than 0.001, the ORs meet a high threshold of statistical significance. In Table 3, there is a huge increase in ORs and CIs between pressure points 2 and 3 (OR from 1.15 to 1.59, CI from [1.13 - 1.17] to [1.53 - 1.65]). This also applies when Post Errors is considered (OR from 1.10 to 1.28, CI from [1.06 - 1.15] to [1.19 - 1.37]), and it suggests a significant increase in the likelihood that an unforced error will occur. This matches what is shown in Figure 3a, where the steepest slopes of our lines are between pressure points 2 and 3.

We also visualize several players to bolster our findings. Figure 3b shows the unforced errors probability of 4 players; 2 Grand Slam winners (Novak Djokovic and Roger Federer) and 2 other players (Casper Ruud and Thiemo De Bakker). Based on Figure 3b, when pressure is 1, the probabilities of hitting an unforced error of Djokovic and Federer are even higher than that of Ruud, suggesting that defining high pressure situations to only include 1 scenario is not enough to understand the chokiness of players. As the value of pressure points increase, we clearly see that Djokovic and Federer's probability of hitting an error decreases while Ruud and De Bakker's overall likelihood is much higher. These results help us accept **H2** and conclude that more successful players tend to choke less during high pressure situations, as their overall error probabilities are below other players and their probability trends are against the mean probability trends.



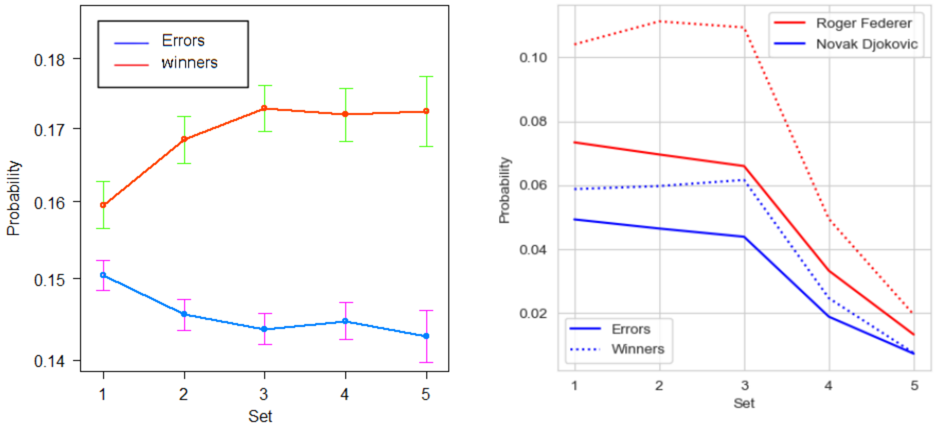
(a) *Mean errors probability comparisons*

(b) *Player examples of errors probability*

Figure 3: *Unforced errors probability based on Pressure Points and Post Error Points. Overall, the mean errors probability is higher when Post Error Points is considered (Figure 3a). A difference in trend is seen between Grand Slam winners (Djokovic, Federer) and others (Ruud, De Bakker), with Grand Slam winners' errors probabilities greatly decrease during very high pressure moments (Figure 3b).*

4.3 Hypothesis 3: The change in probability of unforced errors and winners over time

Figure 4a shows the results of our L_2 and L_3 models - mean probability of players hitting an unforced error and a winner in different sets. According to the figure, the likelihood of unforced errors steadily decreases and the likelihood of winners increase as players approach later sets. According to Tables 4 and 5, the p-values of all predictors are less than 0.0001, suggesting that the effects of set number is statistically significant. Moreover, the predictors' confidence intervals associated with their odd ratios are relatively small, suggesting high precision of the ORs. The values of odd ratios of the predictors are also very close to 1. For winners, they are greater than 1 (between 1.05 and 1.07 for all predictors), meaning that the winners probability is higher as the set number increases, whereas those



(a) *Mean errors and winners probability over set* **(b)** *Player examples of errors and winners probability over set*

Figure 4: *Unforced errors and winners probability based on set number. The mean probability of unforced errors steadily decreases, while that of winners increases over set (Figure 4a). A difference in trend is seen between the probability of winners and unforced errors of player examples, where both the winners (dotted) and unforced errors (bold) probability greatly decreases after set 3 (Figure 4b).*

values are less than 1 for unforced errors (between 0.94 and 0.96 for all predictors), suggesting the opposite.

We visualize the errors and winners rate of 2 example players who are Grand Slam winners: Roger Federer and Novak Djokovic. According to Figure 4b, the trends in errors of these 2 players follow the mean trend in Figure 4a, but at a lower overall rate. However, the trend of winners rate are quite the opposite of the mean trend after set 3, and they are also lower than the overall rate. One explanation is that these players usually win the match in 3 sets and they play sets 4 and 5 relatively rarely. These longer matches may continue to that length because they are against stronger opponents, who make it harder to hit winners against them.

The results we found oppose our hypothesis. We initially expected an increase in unforced errors and a decrease in winners. One potential explanation to these results is the exclusion of *forced errors* in our analysis. Unlike unforced errors, forced errors are errors that players make when a ball is too difficult to return. Forced errors make up for a large proportion of the points in a match. Psychologically, a player causing their opponent to make a forced error can be seen as similar to making a winner, but since the opponent touches the ball before the point ends, it cannot be categorized as a winner. The reduced rate of winners in later sets, seen in Figure 4b, could be explained by fatigued players catching up to the ball less often later in a match. In conclusion, we reject our **H3** and conclude that the rate of winners increases and the rate of unforced errors decrease over the course of a match.

5 Discussion

5.1 Interpretation

In **H1**, we explain the definition of streakiness in tennis and our results show players do not predictably perform better after winning many recent points, nor worse after losing. By analyzing gameplay changes between sets through **H3**, we learn that players hit more winners and less errors as a tennis match goes on. Moreover, from the results of **H2** we know that higher levels of pressure make players more likely to hit errors in the next point. Developing the mental fortitude to win these high pressure points, which are typically more pivotal than low-pressure points, is vital to winning a match. Figures 3b and 4b tell us that the most successful players perform better under high-pressure conditions than other tour players. Our analysis suggests they have a better mentality during high pressure moments and longer matches, which leads to their greater achievements. Therefore, other players who want to be more successful in their careers may want to focus on maintaining their mental poise when dealing with fatigue and pressure.

DJOKOVIC 2 7 15
FEDERER 2 8 40
Nadal 6 6 1 30
Medvedev 7 3 5 40

(a) *Wimbledon Final 2019* **(b)** *Nitto ATP Finals Round Robin 2019*

Figure 5: *Examples of choking moments. In Figure 5a, Djokovic saved Federer's 2 match points and went on to win the final set tiebreak 13 - 12 and won the Wimbledon 2019. In Figure 5b, Nadal saved Medvedev's match point and went on to win the final set tiebreak 7 - 6.*

The importance of avoiding choking can be demonstrated from some key moments in recent tournaments. A well-known example is the final set of Wimbledon Final 2019 [14], where Roger Federer had 2 championship points against Novak Djokovic. Based on Figure 5a, for Djokovic, these 2 points are considered level 4 pressure (late in the game, final set, opponent one game away from winning, facing game point). However, Federer made 2 unforced errors, allowing Djokovic to break back and win the final set tiebreaker 13 - 12 and Wimbledon 2019. Later that year, Rafael Nadal made an even more unbelieveable comeback from 5 - 1 down, facing match point against Daniil Medvedev in Nitto ATP Finals Round Robin [13]. Based on Figure 5b, this situation is considered a level 5 pressure (late in the game, final set, opponent one game away from winning, facing game point and break point) for Nadal. However, Medvedev hit many unforced errors and eventually lost the final set tiebreaker 7 - 6. Since these moments, Federer has not beaten Djokovic again until his recent retirement, and Medvedev has lost twice more against Nadal despite his higher ranking. These are some examples of how a player's mentality can greatly influence their performance on key points and even the match's result. Tennis players seeking to come out on the right side of these close matches would do well to work on their mentality in high-pressure situations like the ones Federer and Medvedev faced. They would also do well to work on recovering mentally after errors, even though unforced errors can be mentally deflating.

5.2 Validity

For **H1**, our construction of streakiness is defined to capture streakiness on short time scales but not long ones, such as across matches. So some types of streaks that spectators might care about are not included. For **H2**, our initial construct validity only includes break points as the sole high pressure situations. Break points are typically more pivotal to the match's result, but they are not the only types of points that might be high-pressure. Therefore, our later analysis includes more factors that we hypothesize would make a point high-pressure based on our knowledge of tennis. For **H3**, we have good construct validity because the variables we claim to measure are exactly those recorded in our dataset.

In **H2**, while we initially found no evidence that the opponents a player faced influenced a player's choking percentage, our data may be confounded by weaker players more consistently finding themselves in high-pressure situations. In **H3**, there may be some selection bias in later sets because not all matches are the same length. However, professional tennis matches are all at least two sets long and our trends hold between set 1 and 2, so the effect on our results is likely minor.

Finally, we consider external validity. Our data comes solely from Grand Slam tennis matches, so our analysis is applicable for professional tennis players, but the results may not be applicable to the general tennis audience. Moreover, tennis gameplay may change over time, so past or future data may not follow the same trends.

5.3 Limitations

There are several limitations to our analysis that could have led us to our current results. One notable limitation is our datasets. For example, we only have complete Grand Slam data from 2011 to 2021, data on older players such as Roger Federer does not include their prime playing years. Also, data collecting systems of several Grand Slam tournaments changed in 2018, leading to missing features and inconsistencies among the names and quantities of features between years. Moreover, the Match Charting dataset does not contain every match, but only a small subset of matches contributed voluntarily by anonymous users, which may cause selection effects. We are also missing point-by-point data of tournaments besides Grand Slam, so phenomena such as "one-tourmanent wonder" players in Grand Slams may lead us to wrong conclusions. When attempting to analyze factors affecting a player's mentality, one shortcoming of our approach is that we cannot accommodate for internal and external factors, such as climate, court atmosphere, crowd, and injuries. This data is not recorded in any tennis datasets available to us.

5.4 Future Works

In **H1**, while we concluded that winning or losing multiple points previously does not have a great impact on winning the next point, in **H2**, an indication of whether a previous point is an unforced error or not does have an impact on the chance of winning the next point. These two results are somewhat contradictory. A compelling avenue for future work would be to attempt to resolve these findings.

Another compelling future study could examine whether streakiness exists on larger time scales, such as between sets or between matches.

6 Related Works

Prior to our research, there have been several analyses on similar topics in the game of tennis.

Klaassen and Magnus [7] investigated if tennis points are independently and identically distributed using a linear model, where a player's performance on a point was a binary variable. In contrast, our analysis on streakiness uses linear and logistic regression models with inputs of win percentages and the outcomes of recent points. Their test results strongly rejected the hypothesis that men's singles points are iid, with a p-value of 0.0003, implying the presence of streakiness. The authors also found that streakiness was more pronounced among weaker players, with a p-value of 0.022.

Cohen-Zada et al. [3] investigated gender differences in choking during tennis matches. Using data from the 2010 Grand Slam tournaments, the results showed significant evidence ($p < 0.01$) that men choke under pressure. The results for women are mixed, but generally, women tend to choke 50% less than men. These results may be linked to the difference in cortisol levels between men and women. This tells us that choking exists, and the level of choking can vary significantly between people.

Maquirriain et al. [8] investigated the effect of fatigue on gameplay for men during the Wimbledon 2015 tournament. The overall serve speed during the first set (177.07 ± 10.28 km/h) and fifth set (176.11 ± 11.74 km/h) were not significantly different (p-value = 0.34). Similarly, serve accuracy as measured by percentage of double faults did not vary significantly (p-value = 0.974) between the first (2.8 ± 3.0) and fifth (2.8 ± 3.4) set. This shows that gameplay does not change significantly in Men's tennis matches, as opposed to our results.

7 Conclusion

In this paper, we studied multiple psychological factors affecting a player's performance and thus a match's result. We used different statistical analysis techniques and machine learning models to validate our hypotheses. Our hypotheses and results showed that not all psychological factors affect players—for example, the results of previous points did not substantially affect the probability that a player wins the next point. We did find a positive correlation between the level of pressure of a point and players' tendency make unforced errors, but our approaches did not come to the same conclusion on whether different players choke more than others. If it is possible for tennis players to train their mentality to perform better in high-pressure situations, such training could improve a player's results.

References

[1] Alberto Arcagni, Vincenzo Candila, and Rosanna Grassi. A new model for predicting the winner in tennis based on the eigenvector centrality. *Annals of Operations Research*, March 2022.

[2] Fred Bowen. Tennis star Naomi Osaka revealed mental health struggles many people face. *The Washington Post*, June 2021.

[3] Danny Cohen-Zada, Alex Krumer, Mosi Rosenboim, and Offer Moshe Shapir. Choking under pressure and gender: Evidence from professional tennis. *Journal of Economic Psychology*, 61:176–190, 2017.

[4] Alexander De Seranno, Toon De Pessemier, and Luc Martens. Predicting Tennis Matches Using Machine Learning, 2020.

[5] Matias Grez. 'It was so dark': Tennis star Nick Kyrgios opens up on mental health struggles. *CNN*, May 2022.

[6] David J. Harris, Samuel J. Vine, Michael W. Eysenck, and Mark R. Wilson. Psychological pressure and compounded errors during elite-level tennis. *Psychology of Sport and Exercise*, 56:101987, 2021.

[7] Franc J. G. M. Klaassen and Jan R. Magnus. Are Points in Tennis Independent and Identically Distributed? Evidence from a Dynamic Binary Panel Data Model. *Journal of the American Statistical Association*, 96(454):500–509, 2001.

[8] Javier Maquirriain, Roberto Baglione, and Marcelo Cardey. Male professional tennis players maintain constant serve speed and accuracy over long matches on grass courts. *European Journal of Sport Science*, 16(7):845–849, 2016. PMID: 26960753.

[9] Jeff Sackmann. ATP Tennis Rankings, Results, and Stats. *GitHub repository*, 2015.

[10] Jeff Sackmann. Grand Slam Point-by-Point Data, 2011-present. *GitHub repository*, 2015.

[11] Jeff Sackmann. The Match Charting Project. *GitHub repository*, 2015.

[12] Cindy Shmerler. The Mental Health of Tennis Players Is No Longer in the Shadows. *The New York Times*, January 2022.

[13] ATP Staff. Nadal Survives To Complete Epic London Comeback. *ATP Tour*, November 2019.

[14] Ravi Ubha. Djokovic saves match points, beats Federer in historic Wimbledon final. *CNN*, July 2019.

[15] Ting Wang, Benjamin Graves, Yves Rosseel, and Edgar C. Merkle. Computation and application of generalized linear mixed model derivatives using lme4. *Psychometrika*, 87(3):1173–1193, feb 2022.

[16] Jack C Yue, Elizabeth P Chou, Ming-Hui Hsieh, and Li-Chen Hsiao. A study of forecasting tennis matches via the Glicko model. *PLoS One*, 17(4), April 2022.

A Appendix

Table 3 summarizes the statistical outcomes of our Model L_1, modelling the effects of pressure points and post error points on unforced errors probability. Each predictor is a combination of 2 independent variable, Pressure Point p and Post Error Point q, with $p \in [0, 4]$ and $q \in [0, 1]$.

Predictors	***Odds Ratio (OR)***	***Confidence Interval (CI)***	***p-value***
Intercept	0.17	0.17 - 0.17	<0.001
Pressure 1	1.01	1.00 - 1.02	0.238
Pressure 2	1.15	1.13 - 1.17	<0.001
Pressure 3	1.59	1.53 - 1.65	<0.001
Pressure 4	1.71	1.55 - 1.88	<0.001
Post Error	1.22	1.20 - 1.24	<0.001
Pressure 1 * Post Error	1.06	1.03 - 1.09	<0.001
Pressure 2 * Post Error	1.10	1.06 - 1.15	<0.001
Pressure 3 * Post Error	1.28	1.19 - 1.37	<0.001
Pressure 4 * Post Error	1.26	1.06 - 1.49	0.009

Table 3: *Odds Ratio, Confidence Interval and p-values of Model* L_1.

Table 4 summarizes the statistical outcomes of our Model L_2, modelling the effects of set number on unforced errors probability.

Predictors	***Odds Ratio (OR)***	***Confidence Interval (CI)***	***p-value***
Intercept (Set 1)	0.18	0.17 - 0.18	<0.001
Set 2	0.96	0.95 - 0.98	<0.001
Set 3	0.95	0.93 - 0.96	<0.001
Set 4	0.96	0.94 - 0.97	<0.001
Set 5	0.94	0.92 - 0.96	<0.001

Table 4: *Odds Ratio, Confidence Interval and p-values of Model* L_2.

Table 5 summarizes the statistical outcomes of our Model L_3, modelling the effects of set number on winners probability.

Predictors	***Odds Ratio (OR)***	***Confidence Interval (CI)***	***p-value***
Intercept (Set 1)	0.19	0.19 - 0.19	<0.001
Set 2	1.05	1.04 - 1.06	<0.001
Set 3	1.07	1.06 - 1.09	<0.001
Set 4	1.07	1.05 - 1.09	<0.001
Set 5	1.07	1.05 - 1.10	<0.001

Table 5: *Odds Ratio, Confidence Interval and p-values of Model* L_3.