Predicting the Outcome of Tennis Matches Project Final Report

Xuefeng Li
a1893802

August 17, 2024

Report submitted for **Research Project** at the School of Mathematical Sciences, University of Adelaide

THE UNIVERSITY of ADELAIDE

Project Area: **Predicting Tennis Match**
Project Supervisor: **Dylan Morris**

In submitting this work I am indicating that I have read the University's Academic Integrity Policy. I declare that all material in this assessment is my own work except where there is clear acknowledgement and reference to the work of others.

I give permission for this work to be reproduced and submitted to other academic staff for educational purposes.

OPTIONAL: I give permission this work to be reproduced and provided to future students as an exemplar report.

1 Abstract

This research project is conducted to create and validate various models for predicting tennis match outcomes. The study compares several models, such as the classic Elo model, an opponent-adjusted stochastic performance model, and more advanced techniques like the FiveThirtyEight Elo prediction and Glicko model.

According to the study results, the FiveThirtyEight based on Elo model significantly outperforms baseline models while the more complex model like Glicko model was not performed as well as we expected. To predict outcome and set a benckmark for whether the model is usable, we have also attempted on BCM model which relect betting market statistic of tennis match

2 Introduction

2.1 Predicting Tennis Match

There are many incentives why people kept attempting predict the outcome of tennis match. Firstly, the money incentives, people bet on the match outcome making it a billion-dollar market Statista (2021)/. Secondly As one of the most popular sports in the world, Tennis is played by millions of people on the earth and the sport channels would broadcast grand slams actively and provide match predictions to fascinate their audiencePeters (2017). Thirdly, Off the field, the coach team would analyze all possible outcomes based based on the player's historical data, so that they could tickle possible outcome during the matchKokta (2020).

The increasing availability and sufficiency of data, along with development of mathematical methods, have enabled researchers to construct and train more accurate models for predicting match outcomes, with open and free access to historical match and betting statisticsSackmann (2019); Tennis Data (2024).

To address the need of predicting the outcome of tennis matches more accurately, researchers took a lot of effort on developing and the statistic has been proven effective on predicting tennis match outcome.(Barnett and Clarke, 2002, 2005; Newton and Keller, 2005)

2.2 Objective

Therefore, our objective of this research project is to construct, validate, and calibrate different models to better predict the outcome of tennis match, and also compare the performance of the models to select best performing one.

The research begins with two main approaches, Elo model based on binary win-loss outcome of historical matches and opponent-adjusted model based on the serve statistic in the game.

Although our research started from Elo model and opponent-adjusted model. The current scope of research on predicting tennis match outcome is based on four types of model applied to predict tennis match outcome: point-based model, paired comparison models, regression based modelsKovalchik (2016), and then machine learning models which emerged dramatically in recent years(A. Šarčević and Krajna, 2022), while our research scope focused on the first two: point-based model, paired comparison models, plus a BCM model.

3 Background

3.1 Data

For this research, we investigated the match data from the ATP Tennis Match, the premier tennis tour for men's singles players, organized by the Association of Tennis Professionals (ATP), including well-known Grand Slams such as Australian open. When composing the dataset from the repository constructed by Sackmann (2019). we took some factors into consideration:

- **Data Validity:** Data before 2000 are less complete than more recent data, making post-2000 data more reliable for analysis.
- **Career Length:** According to Tennis Planet (2020), the average career length of ATP players is 9-10 years, which influenced our selection of the time span.
- **Pandemic Impact:** The COVID-19 pandemic caused many disruptions, impacting match results and tournament schedules. Therefore, we narrowed the time span to a 10-year range from 2010 to 2019 to avoid these anomalies.

We selected two representative players, Daniel Evans and Feliciano Lopez for they actively parcitipating matched in the 9 year range we selected.

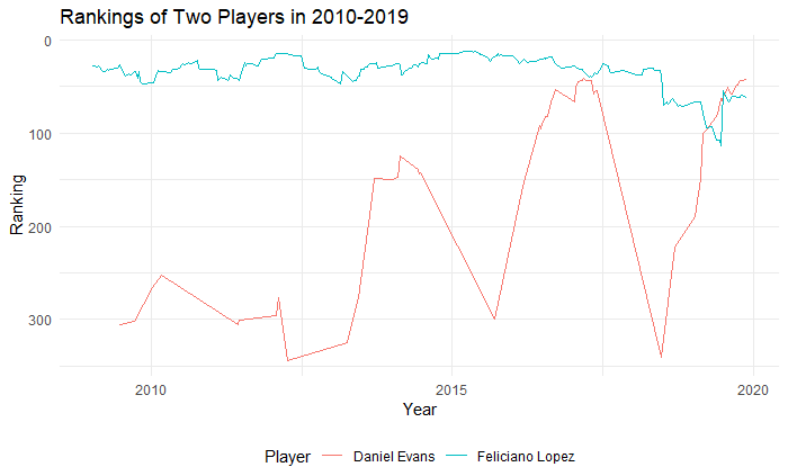


Figure 1: the ranking movements for players' ATP ranking

3.2 EDA

To gain a brief understanding of the dataset, we conducted an Exploratory Data Analysis (EDA). Surface type and tournament level are widely recognized as factors that could impact tennis match outcomes.(McHale and Morton, 2011; S.M. Ma and Ma, 2013; P. Gorgi and Lit, 2019)

3.2.1 Surface Proportion Analysis Across Tournament Levels

As the bar chart shows, hard courts are the predominant surface in Masters 1000 (M) and ATP 500 (A) tournaments, comprising approximately 70.9% and 54.0% of matches, respectively This prevalence underscores the versatility and global availability of hard courts. Clay courts, known for their slower pace, are also significant, making up 33.9% of matches in ATP 500 tournaments and 29.1% in Masters 1000 tournaments.

In contrast, Grass courts are only featured in Grand Slam (G) events, where they account for 25% of the matches, driven primarily by Wimbledon. And The exclusive use of hard courts in the Tour Finals (F) reflects the practical considerations of indoor play. The Davis Cup (D) shows a diverse use of surfaces, with hard courts comprising 61.1% and clay courts 34.8%.

3.2.2 Analysis by Surface

the highest average serve points(194.5) is from Grass courts as its featured in Wimbledon that play best-of-5 sets hence more serves points. on the contrary, despite having a faster bounce, carpert surface had a lower average of 156.2 serve points, likely due to the shorter duration of lower-tier matches where applied carpet more than Grand slams. The fraction of serve points won was highest on Grass courts (approximately 43.28%), proving the well-known advantage for servers on this surface due to the lower bounce and hence quicker points.

3.2.3 Analysis by Tournament Level

Analyzing by tournament level revealed that Grand Slam events (G) had the highest number of matches (15,668) and a large pool of unique players (901). This is expected given the Grand Slam's extended draw sizes and the inclusion of both main draw and qualifying rounds. Conversely, lower-tier events (e.g.Davis cups) had fewer matches and players.

With longer duration of match, highest average serve points played on higher-tier Tournament is higher than lower tier one, while the fraction of serve points won of lower-level tournaments are lower hovering

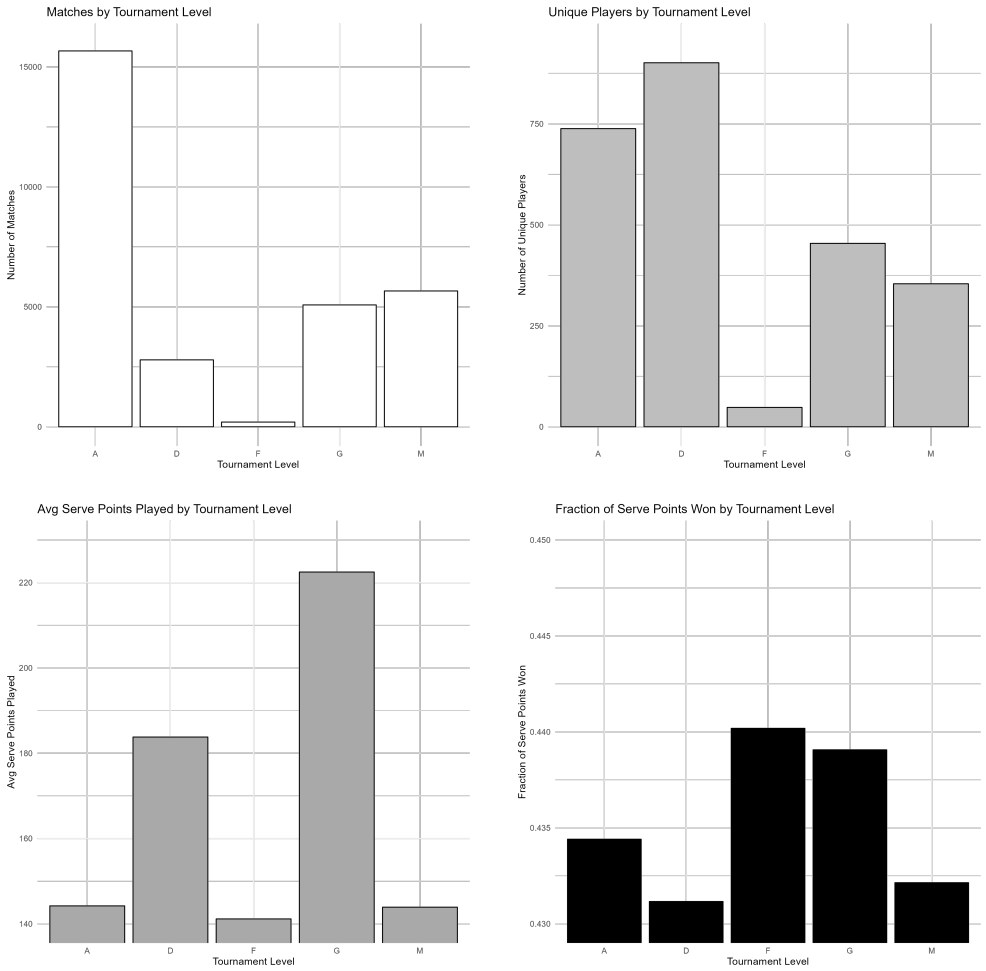


Figure 2: Matches, Player number, Serve Points, and percentage Serve Points won by different tournament levels.

around 43%, but slightly higher in top-tier tournaments, which may reflect the ability of players entering into higher-level tournaments have more capability to return the serve.

overall our findings at EDA corroborate the understanding that both surface type and tournament level significantly influence match dynamics and player participation.

3.3 Development of Prediction Models

The development of tennis match prediction models has progressed through several distinct phases, with the majority of approaches falling into three main categories: hierarchical/point-based models, pairwise comparison models, and regression/machine learning models Klaassen and Magnus

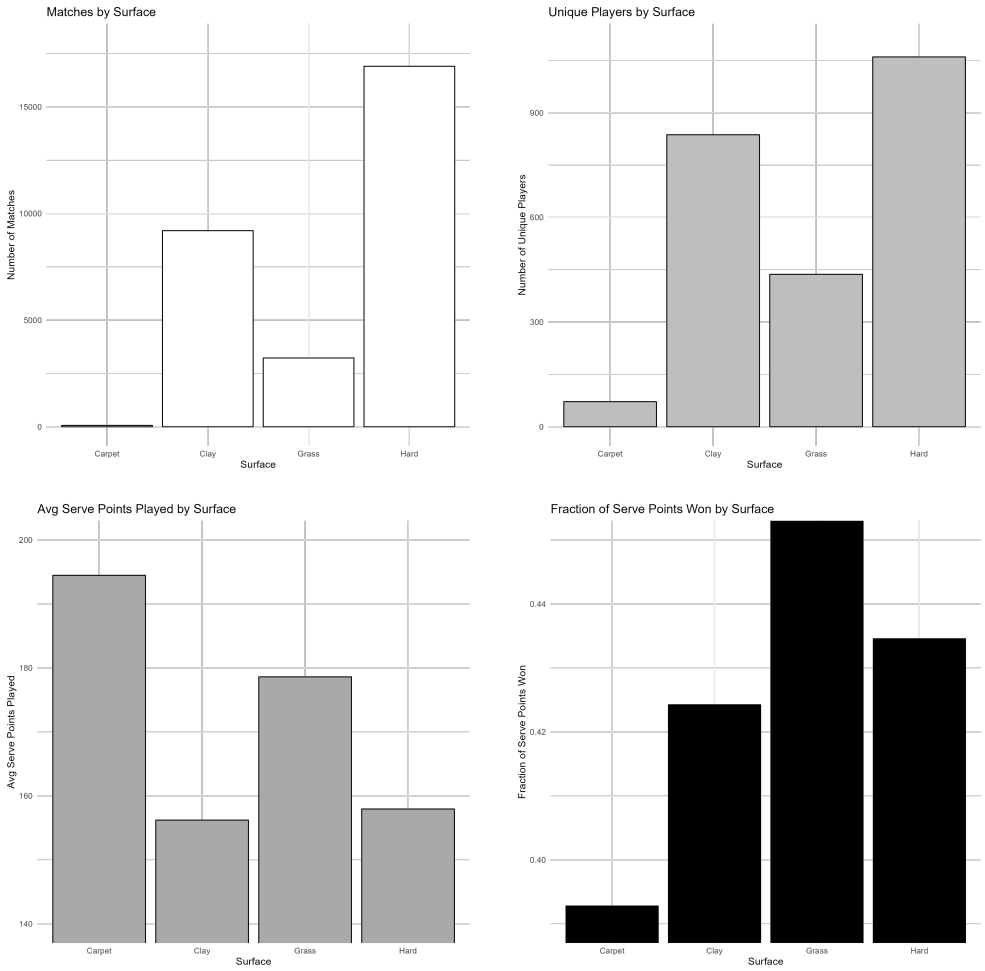


Figure 3: Matches, Player number, Serve Points, and percentage Serve Points won by different surfaces.

(2003); McHale and Morton (2011); Boulier and Stekler (1999).

3.3.1 Point-Based Models

Early work by klaassen(2001)Klaassen and Magnus (2001) and barnett(2002)Barnett and Clarke (2002) laid the groundwork for this approach by validating the i.i.d. assumption and modeling tennis matches as a series of i.i.d. points. These models utilized historical service statistics to calculate point-winning probabilities, which were then aggregated to estimate the likelihood of winning a game, set, or match. (Barnett& Clarke,2005)Barnett and Clarke (2005) further advanced this methodology by introducing recursive formulas that allowed for real-time match predictions, a significant step towards dynamic modeling in tennis.

However, point-based models faced challenges due to biases introduced by varying opponent skill levels, which could distort predictions To address this, knottenbelt(2012)Knottenbelt et al. (2012) developed a common opponent averaging method, which improved the estimation of service probabilities by focusing on matches where both players had faced the same opponents. This refinement helped to mitigate the biases and improved the model's predictive accuracy.

Later, ingram(2019)Ingram (2019) introduced a Bayesian hierarchical model that enhances point-based tennis match predictions by incorporating surface-specific skills and tournament effects, significantly improving prediction accuracy and reducing log loss compared to previous point-based models

3.3.2 Pairwise Comparison Models

Pairwise comparison models, such as the Bradley-Terry model, compare players in pairs to estimate the probability of one player defeating another(bradley,1952)Bradley and Terry (1952). McHale and Morton (2011) applied this model to tennis, using a single skill parameter for each player. The model has been further enhanced by incorporating time-varying components, as demonstrated by Baker and McHale (2014), allowing the model to account for changes in player abilities over time.Adopting this nuanced approach is expected to increase the model's ability to predict performance, reflecting how players' performance impacted by varied surfaces and competitive environments in real life.

The Elo rating system, widely used in sports analytics, operates on a similar principle. However, unlike the fixed-period optimization in the Bradley-Terry model, the Elo system dynamically updates player ratings after each match, as noted by (kovalchik,2016)Kovalchik (2016). This adaptability has made Elo one of the most popular methods for predicting tennis match outcomes.

3.3.3 Regression and Machine Learning Models

Regression models have been widely used in tennis match prediction, particularly in the early stages of development. boulier(1999)Boulier and Stekler (1999) and clarke(2000)Clarke and Dyte (2000) used regression techniques with features such as ranking points and seeding to estimate the probability of a player winning a match. More recent studies, like those by sipko(2015)Sipko and Knottenbelt (2015), have expanded the feature set to include detailed match statistics such as serve percentages and aces, often calculated using point-based methods.

Neural networks, a form of machine learning, have also been applied to tennis match prediction. Early work by somboonphokkaphan(2009)A. Somboonphokkaphan and Lursinsap (2009) demonstrated the potential of neural networks to improve predictive accuracy by learning complex patterns in historical match data. However, as wilkens(2020)Wilkens (2020) notes, while machine learning models can be powerful, they often require large datasets and may still struggle to outperform simpler models like Elo when the most relevant features are well-captured by these traditional approaches.

4 Methods

In perspective of mathematical, tennis fits modelling well for its nature. Unlike team-playing games like basketball or football. In a single tennis match, only two players involved and there would and only could be a binary outcome, one player win and the other one lose. The Matches process can be easily modelled as discrete stochastic processes. Tennis focuses on the score instead of time and the probability of scoring for each serve has been proven to follow assume an independent and identical point distribution. Moreover, there is a comprehensive database for the tennis match, for example the ATP tennis match database.(Sackmann, 2019)

4.1 Baseline Model

4.1.1 Naive Model

For this research, to establish a baseline for model performance, we first employed a very basic model to provide an intuitive benchmark, simplying predicting the player with higher ATP ranking would win the lower one

4.1.2 Logistic Model

Next, we applied a simple logistic model to predict the probability of the higher-ranked player defeating the lower-ranked one. formula as follows

$$P(x) = \frac{1}{1 + e^{-(\beta_1 x)}}$$

Where:

- P(higher_rank_won) is the predicted probability of the outcome
- e is the base of the natural logarithm

- β_1 is the model coefficent that could be adjusted for the predictor x
- x is the ranking difference between players

Our logistic model plot is as follows: the closer we get to the top left, the more likely it is that a higher-ranked player wins. However, when the difference is not significant, it would be less likely tell which side of the match will win.

4.2 Elo Model

The Elo model is widely applied here to predict the outcomes of tennis matches including current ATP ranking points calculation. As we discussed above tennis match prediction naturally suites Elo model as The Elo model was originally developed for another two-player zero-sum games, the chess. In the Elo model, the player would have a corresponding change after each match they played based on the outcome of match and their opponent's ranking. As the Elo model could utilize historical data while naive model and logistic could only rely on the static ATP ranking at the time of the match. The Elo model is supposed to outperform the baseline models on accuracy and validity. Following the previous literature by Topspin (2019); G. Angelini and Angelis (2022), we set the initial Elo score to 1500. Following the classic Elo model, we introduced three improved model, the K-factor model, fivethirtyeight model and Glicko Model.

4.2.1 K factor model

In the K-factor version of the Elo model, a constant K determines the magnitude of the rating adjustment after each match. The change in a player's rating is calculated using the formula:

$$\Delta\text{Rating}_i = k \times (\text{Match Outcome}_i - \text{Predicted Outcome}_i) \quad (1)$$

At the beginning, we set the factor k as 25, but after tuned k factor based on the performance metrics calculated. we moved it down to 14 as optimal factor, which consistent with previous research by jonas (2016), and the match outcome is a binary outcome that same with the two models above, whether the player with higher rating won, if the outcome is the opposite to the prediction, then the model would be adjusted according to better fit the reality. and then we update the players with the new rating which are the sum of the change in the Elo rating and

their previous rating. Similar to the ranking, the rating of a player could be adjusted up or down. However, a limitation of the traditional Elo model is that it doesn't account for how long the players played for, or how many matches they had attended already. For new atheltes getting into the field, we should adjust their ranking more aggressively based on they early matches outcomes so that they could locate a ranking position playing with compatible opponents. This issue is highlighted by when they debating about whether Serena Williams, one of the most famous female tennis player, is the Greatest of all Time(GOAT) player (Morris and Bialik, 2017).

4.2.2 FiveThirtyEight model

To address the issue that the traditional Elo model is unable to tackle, we also applied the model introduced by Morris(2017)Morris and Bialik (2017), using δ, ν, and σ to form a dynamic k-factor as formula showing below.

$$K_i(t) = \frac{\delta}{(m_i(t) + \nu)^\sigma}$$

Here, m represents the number of matches a player has competed in during the specified period, while δ, ν and are parameters that can be tuned while σ can stand constant as it stands for a mutiplier role as δ. This adaptation allows the model to account for player activity and time they have played, adjusting their ranking more properly.

4.2.3 Glicko model

On top of the Elo model, the Glicko model was constructed by Mark Glickman (1999)Glickman (1999). The most significant difference in the Glicko model is that each player is assigned a value for rating deviation (RD), where RD is a factor taking the uncertainty of the rating into account.

After competing in matches, these values are updated based on the outcomes against their opponents. The updated rating r' is calculated using the following formula:

$$r' = r + \frac{q}{\frac{1}{RD^2} + \frac{1}{d^2}} \sum_{j=1}^{m} g(RD_j)(s_j - E(s|r, r_j, RD_j))$$

Here, $q = \ln(10)/400 = 0.0057565$, and $g(RD_j)$ is a scaling function defined as:

$$g(RD_j) = \frac{1}{\sqrt{1 + \frac{3q^2 RD_j^2}{\pi^2}}}$$

The expected win probability $E(s|r, r_j, RD_j)$ against an opponent with rating r_j and RD RD_j is calculated as:

$$E(s|r, r_j, RD_j) = \frac{1}{1 + 10^{-g(RD_j)(r-r_j)/400}}$$

The term d^2 is the variance of the expected outcome, calculated as:

$$d^2 = \left(q^2 \sum_{j=1}^{m} g(RD_j)^2 E(s|r, r_j, RD_j)(1 - E(s|r, r_j, RD_j)) \right)^{-1}$$

After calculating the new rating, the updated RD RD' is determined by:

$$RD' = \sqrt{\left(\frac{1}{RD^2} + \frac{1}{d^2} \right)^{-1}}$$

Overall, even though the glicko model still using ranking update and comparison to estimate winning prbability and outcome, the mathematical design is much more delicated than Elo rating system

4.2.4 TrueSkill Model

The TrueSkill model, developed by Microsoft, is a Bayesian pair comparison rating system that significantly improves upon Elo model in competitive games. Like all other improved model mentioned. TrueSkill assume the distribution of players' skill level is a normal distribution, characterized by a mean (μ) and a variance (σ^2). After each match, the model would updates both μ and σ^2 based on the match outcome for further match outcome prediction(ralf et al, 2006)Herbrich et al. (2006).

In TrueSkill, the skill level of player i modeled as $p_i \sim \mathcal{N}(s_i, \beta^2)$, where s_i is the skill and β^2 is the variance capturing performance randomness. The probability that player i wins player j is given by:

$$P(p_i > p_j) = \Phi \left(\frac{\mu_i - \mu_j}{\sqrt{2}\beta} \right),$$

where Φ is the cumulative distribution function of the standard normal distribution. This model's ability to rapidly refine skill estimates through Bayesian updates makes it well-suited for dynamic environments, such as tennis, where player performance can fluctuate based on various factors. While TrueSkill's multiplayer capabilities are not relevant to singles tennis, its adaptability and probabilistic updates provide a more accurate ranking framework than Elo, especially handling team games.

4.3 Point-based model

4.3.1 Scoring system

A typical tennis match scoring system lies follow, beginning with the first serve and progressing through a series of points, games, and sets. If both players reach 6-6 in a set, a tiebreak is introduced to decide the winner of that set, where the first player to reach 7 points with a 2-point lead wins the set.

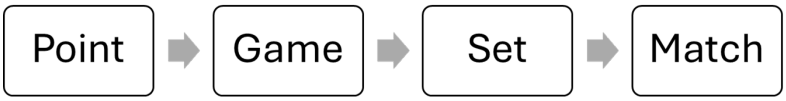


Figure 4: scoring level of tennis match

- **Points**: In tennis, the lowest scoring unit starts from love (no points) and continues as 15, 30, and then to 40 (equivalent to 1, 2, and 3 points, respectively). The game typically ends after reaching 40 because a player must score two points after reaching 40 to win. A tie at 40-40 is called deuce; the server can choose which side to serve from and must win two consecutive points to take the game.
- **Games**: A game is won by the first player to attain 4 points, with at least a two-point margin. If the score is tied at deuce, the player needs to win two consecutive points to secure the game.
- **Sets**: A player wins a set by being the first to reach 6 games, with at least a two-game lead over the opponent. If both players reach 6-6, a tiebreak is played. In the tiebreak, the first player to win 7 points, with at least a two-point lead, wins the set.
- **Match**: A match is a best-of-three or best-of-five set competition. The first player to win the majority of sets wins the match.

4.3.2 IID Assumption

For most point-based models, the IID assumption is essential for the models to be valid. The point-level assumption is examined first. firstly liu et al(2001)Liu (2001) worked on the association between two consecutive points. Klaassen and Magnus klaassen(2001)Klaassen and Magnus (2001) modeled the probability of winning a point during a service and demonstrated that points are neither independent nor identically distributed. In fact, the probability of winning a point is not always constant

barnett(2006)Barnett et al. (2006). Winning a previous point increases the likelihood of winning the subsequent one, consistent with the "hot hand" phenomenon in real matches, where a player experiences continued success after a string of successes. Additionally, the difficulty for the server to win a point is greater during "important" points compared to less significant ones. These effects are more pronounced for weaker players. However, the deviations from the assumption of independence and identical distribution (i.i.d.) are minimal, meaning that in many situations, the i.i.d. hypothesis still serves as a reasonable assumption for simulation and prediction . Following this, a proof of mathematical deduction is provided to further develop the i.i.d. assumption to the game, set, and match levels (newton,2005)Newton and Keller (2005), which provide a substantial foundation for point-based models.

4.3.3 Markov Chain Process

Although the independence and identical distribution assumption is not introduced and proved, the Markov chain and conditional probability have been applied to predict game outcomes in Excel (barnett,2002)Barnett and Clarke (2002), forming the prototype of the point-based model.

The model starts by assuming a single game between two players, A and B, where player A serves with a constant probability p of winning a point. The model uses a Markov chain that represents the game state as the current score in points (for example, a score of 40-15 is represented as 3-1). With a probability p, the state advances from (a, b) to $(a+1, b)$, and with a probability $1-p$, it changes from (a, b) to $(a, b+1)$. Consequently, if $P(a, b)$ denotes the probability that player A wins when the score is (a, b), then the following equation holds:

$$P(a, b) = pP(a + 1, b) + (1 - p)P(a, b + 1)$$

The conditional probability of a specific serve is as follows:

$$P(a, b) = 1 \text{ if } a = 4,\ b \leq 2, \quad P(a, b) = 0 \text{ if } b = 4,\ a \leq 2$$

Consequently, the probability of winning from deuce is given by:

$$\frac{p^2}{p^2 + (1 - p)^2}$$

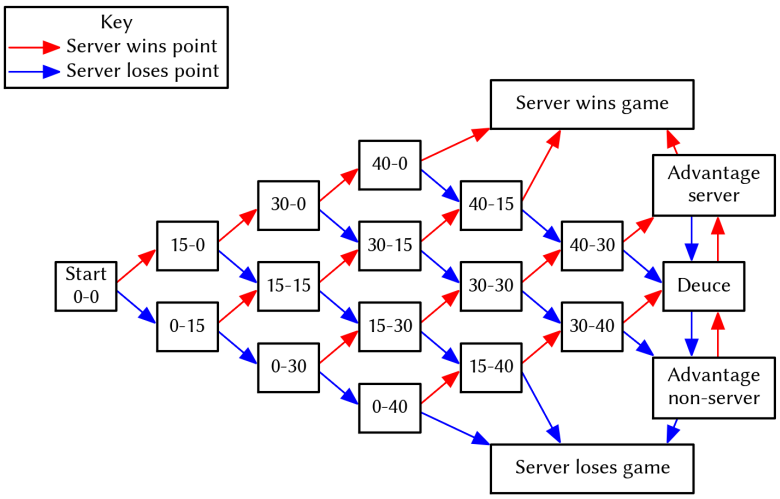


Figure 5: Markov chain scoring process for a game in tennis match

4.4 BCM Model

Betting in sport is widely popular among fans. For a single game, match, or tournament, bookmakers set odds for different outcomes of a sporting event to allow people to bet on their expected outcome's odds.

If the odds for outcomes are $O_1, O_2, \ldots, O_n$, the implied probabilities are:

$$P_i = \frac{1}{O_i}$$

The overround is the amount by which the sum of the implied probabilities of all possible outcomes exceeds 100%, which represents the bookmaker's expected profit for the set.

The total overround for a single event is:

$$\text{Overround} = \sum_{i=1}^{n} \frac{1}{O_i} - 1$$

For example, if the odds for a match between two teams are 1.25 and 3.00, the implied probabilities are 0.50 and 0.33, respectively. The overround would be:

$$\text{Overround} = 0.80 + 0.33 = 1.13 \text{ or } 113\%$$

In this case, the bookmaker has a profit margin of 13%. This is where the overround gets a little more tricky when you have a composed bet, such as accumulators or parlays. When probabilities are multiplied, the total overround of multiple bets is not just a sum of individual overrounds. A

bookmaker's profit margin compounds as the number of selections added to a bet increases. To summarize, the overround in multiples doesn't completely eliminate the potential for a positive return on investment (ROI) from high-odds betting combinations, primarily because of this higher default theoretical expected return (or negative margin) on single selections.

5 Results

5.1 Validation

5.1.1 Accuracy

Accuracy is one of the most fundamental evaluation metrics, providing a straightforward assessment of a model's performance. A high accuracy in predicting the testing dataset indicates that the model performs well.

$$\alpha = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{f(i)=y_i\}} \tag{2}$$

Where:

- α represents the model's accuracy.
- N denotes the number of observations in the validation set.
- $f(i)$ returns the predicted winner of the match.
- $\mathbf{1}$ is a binary indicator showing whether the higher-ranked player wins.

5.1.2 Log-loss

Logistic loss (Log-loss) evaluates the model's classification accuracy by penalizing incorrect predictions. It depends on the predicted probabilities and the actual match outcomes. Lower Log-loss values indicate better model performance. The calculation is given by:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)] \tag{3}$$

Where π_i is the predicted probability that the higher-ranked player wins match i.

5.1.3 Calibration

Calibration assesses the alignment between predicted probabilities and actual outcomes. If the predicted probabilities closely match the observed probabilities, the model is well-calibrated.

$$C = \frac{1}{N}\sum_{i=1}^{N} \pi_i \tag{4}$$

Here, W represents the total number of matches won by the higher-ranked player. A well-calibrated model will have a calibration value approximately equal to 1. If the value exceeds 1, the model overestimates the probability of the higher-ranked player winning; if it is below 1, the model underestimates it.

5.2 Tunning & Analysis

To optimize the model performance, we had run model with different factor value, In context of Elo model, The k factor in the Elo rating system determines the sensitivity of rating updates after each match, the optimal k value should update the player's in a opportune pace changing player's ranking that most accurately reflect the player's ability in elo ranking through the matches

To perform the tuning, we tested a range of k values, from 1 to 50. The metrics we use to reflect model's performance in one value calculated as:

$$\text{Performance Metric} = \text{Accuracy} - \text{Log Loss}$$

This metric balances the trade-off between the model's ability to predict the correct outcome (accuracy) and its confidence in those predictions (log loss).

The resulting plot from the tuning process illustrate how the performance metric changes as the k factor increases. .

From the plot, we observe that the performance initially increases as k increases, reaching a peak at 10. Beyond this point, the performance metric begins to decline, indicating that overly high k values may caused overfitting, where the model becomes too sensitive to individual match outcomes.

5.3 Naive Model

We split the dataset from 2017-01-01, the training set was data before that date and the testing data is the data afterwards.

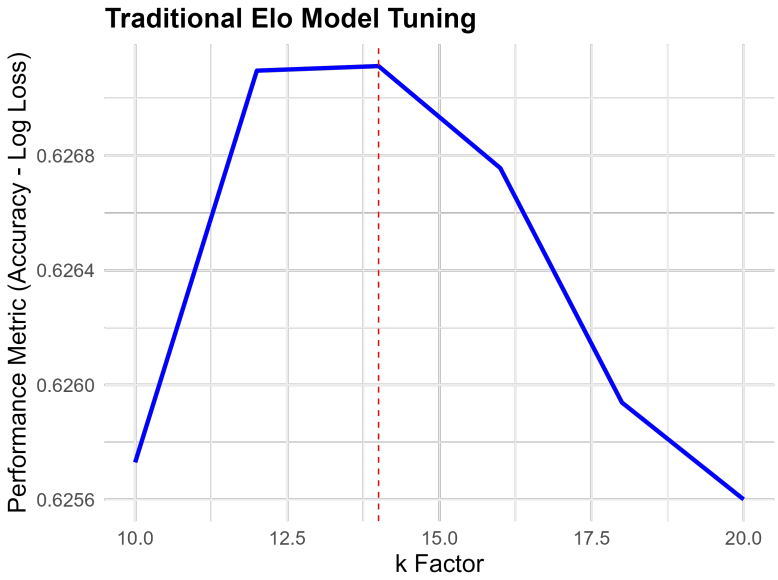


Figure 6: Tuning plot for the traditional Elo model. The red dashed line indicates the optimal k value.

Table 1: Naive Model Validation Statistics

Model	Pred_acc	Log_loss	Calibration	Dataset
naive	0.6780036	0.6291074	1.000000	training
naive	0.6365444	0.6546917	1.000000	testing
naive	0.6645604	0.6372510	1.000000	Full Set

as the table shows, the naive model could have a 63.7 percent accuracy on predicting, and the log-loss is relatively high at 0.655 yet the calibration is exactly 1 which mean the model doesn't have calibration issue. overall, the naive model is subject to moderate over-fitting as the validation statistic of testing model unperformed the overall set.

5.4 Logistic Model

Using the same data split, the result of Logistic Model shows below.

Table 2: Logistic Model Validation Statistics

Model	Pred_acc	Log_loss	Calibration	Dataset
logistic	0.6763549	0.6179265	0.9167392	training
logistic	0.6371239	0.6437330	0.9564547	testing
logistic	0.6661654	0.6246487	0.9265667	Full Set

as the table shows the logistic model could have a 67.7 percent accuracy on predicting, roughly same to naive model. However, the log-loss is relatively higher at 0.775 for testing set, which is also about the same to naive model. yet the calibration is under 1 which mean the model underestimate the result. Besides, the logistic model also has over-fitting issue as the validation statistic of testing model unperformed the overall set.

5.5 K factor Model

The plot in Figure 7 illustrates the performance metric of the Traditional Elo model across different values of the k-factor. The tuning process aimed to determine the k-factor that optimizes model performance, which is measured by a combination of prediction accuracy and log loss.

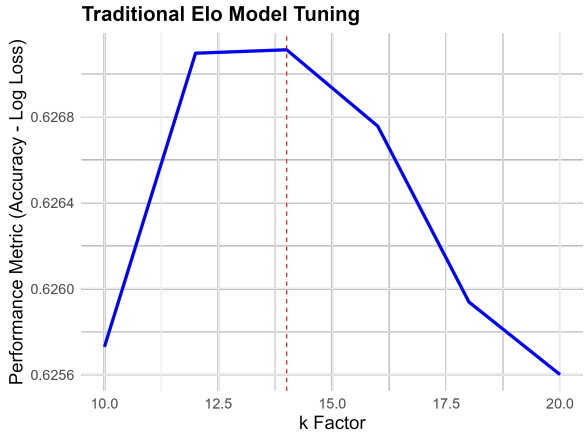


Figure 7: Tuning the k-factor for the Traditional Elo Model. The red dashed line indicates the optimal k-factor.

Analysis:

- The performance metric increases as the k-factor rises, reaching a peak around a k-factor of 12.5 to 15.

- After this peak, the performance metric declines, indicating that higher k-factors may lead to overfitting or reduced model stability.
- The optimal k-factor, around 15, represents the best balance between accuracy and prediction confidence, making it the preferred choice for final model configuration.

Conclusion: The tuning analysis suggests that a k-factor of 15 should be used in the Traditional Elo model to achieve maximum performance. This value optimizes the model's ability to make accurate and confident predictions.

5.5.1 Analysis of Validation Results

Table 3: Elo Model Validation Statistics

Model	Pred_acc	Log_loss	Calibration	Dataset
Elo	0.668	0.417	0.874	Training
Elo	0.614	0.431	0.911	Testing
Elo	0.663	0.418	0.877	Full Set

The validation results for the Elo model are summarized in Table. The Elo model's performance is evaluated based on prediction accuracy (Pred_acc), log loss, and calibration across the training, testing, and full datasets.

- The model achieves a prediction accuracy of 0.668 on the training set and 0.614 on the testing set.
- The log loss values are 0.417 for the training set and 0.431 for the testing set, indicating the model's confidence in its predictions.
- Calibration scores are fairly consistent across datasets, with the full set showing a calibration score of 0.877, indicating that the predicted probabilities are generally reflective of the actual outcomes.

Overall, the Elo model performs reasonably well but shows room for improvement, particularly in prediction accuracy on the testing set.

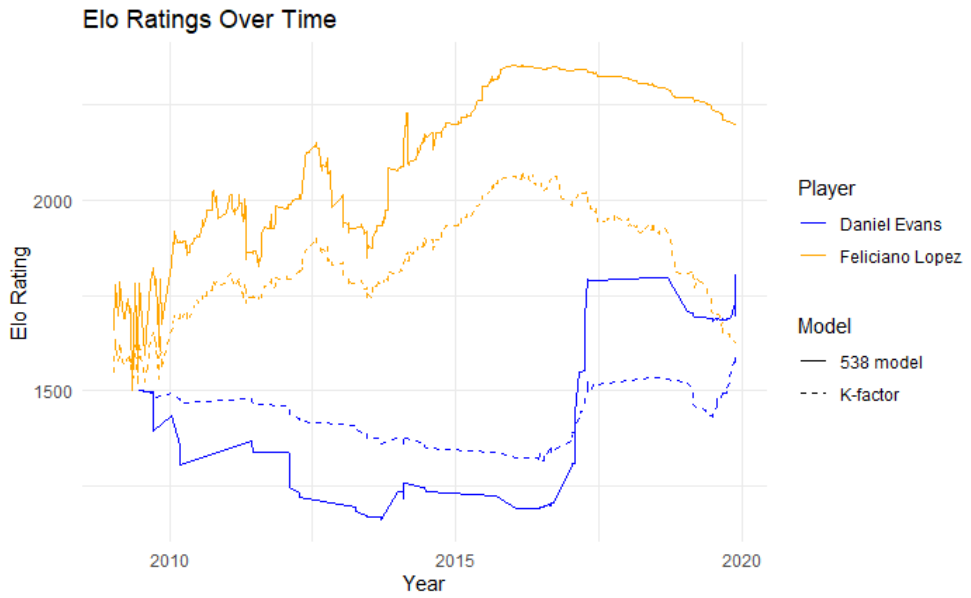


Figure 8: Elo rating movement from 2010 to 2019

5.6 FivethirtyEight Model

5.6.1 Step 1: Initial Grid Search

The initial grid search, visualized in the first heatmap (Figure 9(a)), explores a broad range of values for both ν and δ. The heatmap shows the performance of different combinations, with brighter and lighter colors indicating better model performance. The grid search covered the following ranges:

- δ: 100 to 300, increasing in steps of 50.
- ν: 10 to 30, increasing in steps of 5.

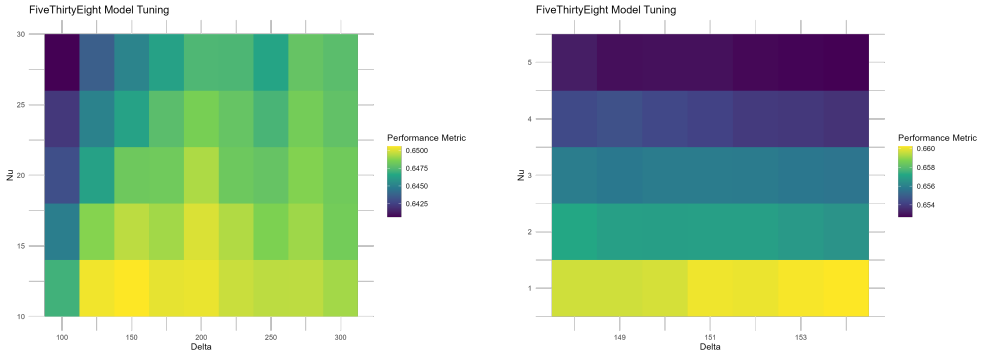
Based on the results from the initial grid search, we observed that the best performance was concentrated in a specific region of the parameter space, particularly where δ values are around 150-200 and ν values are around 10-15.

5.6.2 Step 2: Refined Grid Search

Given the insights from the initial grid search, a more focused and refined search was conducted (Figure 9(b)). This refined grid search narrows down the ranges as follows:

- δ: 148 to 154, increasing in steps of 1.

The tuning process for the FiveThirtyEight model involves a grid search across the parameter space defined by ν (the performance decay factor) and δ (the starting rating value). The goal of this tuning is to identify the combination of ν and δ that maximizes model performance, which is evaluated by a performance metric that balances prediction accuracy and log loss.



(a) Initial Grid Search (b) Refined Grid Search

Figure 9: Tuning analysis of the FiveThirtyEight model parameters ν and δ using grid search. The brighter and lighter colors represent better model performance, with the refined grid search narrowing down the range based on the initial search results.

- ν: 1 to 5, increasing in steps of 1.

The refined heatmap shows a more precise identification of the optimal parameter values. The area of peak performance, represented by the brightest and lightest colors, confirms the findings from the initial search. The best-performing combination from this refined search will be used to finalize the FiveThirtyEight model's parameters

5.6.3 Result Analysis

Table 4: 538 Model Validation Statistics

Model	Pred_acc	Log_loss	Calibration	Dataset
538	0.688	0.392	0.872	Training
538	0.617	0.393	0.914	Testing
538	0.681	0.392	0.875	Full Set

The validation results for the FiveThirtyEight (538) model are summarized in Table 2. The 538 model is evaluated using the same metrics as

the Elo model: prediction accuracy (Pred_acc), log loss, and calibration.

- The 538 model shows strong prediction accuracy, with a training accuracy of 0.688 and a testing accuracy of 0.617, outperforming the Elo model.
- The log loss values for the 538 model are lower than those for the Elo model, at 0.392 for both the training and full sets, and 0.393 for the testing set, indicating more confident and accurate predictions.
- Calibration scores are consistent, with a slight improvement over the Elo model, particularly on the testing set where the calibration score is 0.914.

The 538 model outperforms the traditional Elo model across all metrics, making it a more reliable option for predicting outcomes in this dataset.

5.7 Glicko Model

Due to time limitation, we ran the codes manually and find the optimal c value in our case is 1. and the performance of the

The Glicko model's validation statistics, presented in Table 2, indicate strong performance, particularly in terms of log loss and calibration metrics.

- The Glicko model's prediction accuracy is slightly lower compared to the 538 model, with a training accuracy of 0.641 and a testing accuracy of 0.603.
- However, the Glicko model achieves the lowest log loss values among the models, with a log loss of 0.290 on the training set and 0.305 on the testing set, reflecting its ability to make highly confident predictions that are closer to the true outcomes.
- The Glicko model also exhibits superior calibration, particularly on the testing and full datasets, with calibration scores of 0.939 and 0.928, respectively. This suggests that the Glicko model is more reliable in predicting actual match outcomes.

For the result, the Glicko model's strong performance in log loss and calibration metrics makes it a competitive alternative to both the Elo and 538 models, especially when confidence in predictions is paramount.

6 Discussion

6.1 Subset Analysis

Tournament-level and surface are proven have significant impact on players' performance and hence the match outcome.S.M. Ma and Ma (2013); Corral and Prieto-Rodríguez (2010). Therefore, we had run the EDA and some subset results to address their impact on the match outcome and also the sensitivity of models to the dataset features.

6.1.1 Surface

As shown in EDA, the grass, as mainly featured by Wimbledon Open, are tends to have more serve points and lower fraction of serve point won, hence the match result would be less predictable.

6.1.2 Tournament-Level subset

Instead of Tournament-level, the ATP ranking is more representing the hierarchy among players as high-tier tournaments only give access to top players. The first requirement for players to Grand Slam Tournament is that they need to at least rank around 104 in world ATP ranking at entry time. Therefore, we trim the dataset into Top 50 one and Top 100 one to see how model behave differently.

Here is the calculated performance metrics on testing set.

Table 5: Performance Metrics for Testing Set across Different Datasets

Model	Full Set (All Players)	Top 100 Players	Top 50 Players
Elo	0.614 / 0.431 / 0.911	0.609 / 0.430 / 0.919	0.654 / 0.360 / 0.901
538	0.617 / 0.393 / 0.914	0.609 / 0.316 / 0.947	0.655 / 0.266 / 0.935
Glicko	0.603 / 0.305 / 0.939	0.601 / 0.303 / 0.945	0.655 / 0.271 / 0.919

The table above summarizes the performance of the Elo, 538, and Glicko models across different dataset sizes—Full Set, Top 100 Players, and Top 50 Players—based on testing set metrics. The 538 model shows a significant improvement in accuracy and log loss when the dataset is reduced from all players to the top 50. Accuracy increases from 0.617 (Full Set) to 0.655 (Top 50 Players), and log loss decreases from 0.393 to 0.266, indicating more precise predictions with a smaller, elite group of players.

The Glicko model demonstrates excellent calibration across all datasets, with consistent results ranging from 0.939 (Full Set) to 0.919 (Top 50

Players). This suggests that Glicko's predicted probabilities align closely with actual outcomes, regardless of the dataset size.

The Elo model, trained on fewer players, shows improvements in both accuracy and log loss as well. Accuracy rises from 0.614 (Full Set) to 0.654 (Top 50 Players), and log loss decreases from 0.431 to 0.360, reflecting more certain predictions with fewer outcomes in the reduced dataset.

For the top 50 players, the 538 model slightly outperforms the Elo model in accuracy (0.655 vs. 0.654) and log loss (0.266 vs. 0.360), suggesting more confident predictions from the 538 model. Both models are competitive, with the 538 model holding a slight edge. However, the Glicko model, with a calibration of 0.919, may have a slight advantage in matching predicted probabilities with actual outcomes.

References

A. Somboonphokkaphan, S. P. and Lursinsap, C. (2009). Tennis winner prediction based on time-series history with neural modeling. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, pages 18–20.

A. Šarčević, M. Vranić, D. P. and Krajna, A. (2022). Predictive modeling of tennis matches: a review. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1099–1104.

Baker, R. and McHale, I. (2014). A dynamic paired comparisons model: Who is the greatest tennis player? *European Journal of Operational Research*, 236(2):677–684.

Barnett, T. and Clarke, S. (2002). Using microsoft excel to model a tennis match. In *6th Conference on Mathematics and Computers in Sport*, pages 63–68, Queensland, Australia. Bond University.

Barnett, T. and Clarke, S. R. (2005). Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, 16(2):113–120. Accessed: 2024-04-11.

Barnett, T. J., Brown, A., and Clarke, S. (2006). Developing a tennis model that reflects outcomes of tennis matches. *Conference contribution, Swinburne University.*

Boulier, B. and Stekler, H. (1999). Are sports seedings good predictors?: an evaluation. *International Journal of Forecasting*, 15(1):83–91.

Bradley, R. and Terry, M. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Clarke, S. and Dyte, D. (2000). Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research*, 7(6):585–594.

Corral, J. D. and Prieto-Rodríguez, J. (2010). Are differences in ranks good predictors for grand slam tennis matches? *International Journal of Forecasting*, 26(3):551–563.

G. Angelini, V. C. and Angelis, L. D. (2022). Weighted elo rating for tennis match predictions. *European Journal of Operational Research*, 297(1):120–132.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394.

Herbrich, R., Minka, T., and Graepel, T. (2006). Trueskilltm: A bayesian skill rating system. *Advances in Neural Information Processing Systems*, 20:569–576.

Ingram, M. (2019). A point-based bayesian hierarchical model to predict the outcome of tennis matches. *Journal of Quantitative Analysis in Sports*, 15(4):313–325. Accessed: 2024-08-01.

jonas (2016). Predicting atp tennis match outcomes using serving statistics. Accessed: 2024-08-04.

Klaassen, F. and Magnus, J. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*, 148(2):257–267.

Klaassen, F. J. G. M. and Magnus, J. R. (2001). 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. Accessed: 2024-08-13.

Knottenbelt, W. J., Spanias, D., and Madurska, A. M. (2012). A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers and Mathematics with Applications*, 64(12):3820–3827. Theory and Practice of Stochastic Modeling.

Kokta, M. (2020). Predicting atp tennis match outcomes using serving statistics. Accessed: 2024-07-20.

Kovalchik, S. A. (2016). Searching for the goat of tennis win prediction. *Journal of Quantitative Analysis in Sports*, 12(3):127–138.

Liu, Y. (2001). Random walks in tennis. *Missouri Journal of Mathematical Sciences*, 13(3):154–162.

McHale, I. and Morton, A. (2011). A bradley-terry type model for forecasting tennis match results. *International Journal of Forecasting*, 27(2):619–630.

Morris, B. and Bialik, C. (2017). Serena williams and the difference between all-time great and greatest of all time. Accessed: 2024-04-19.

Newton, P. and Keller, J. (2005). Probability of winning at tennis i. theory and data. *Studies in Applied Mathematics*, 114(3):241–269.

P. Gorgi, S. K. and Lit, R. (2019). The analysis and forecasting of tennis matches by using a high dimensional dynamic model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4):1393–1409.

Peters, J. (2017). Predicting the outcomes of professional tennis matches. Accessed: 2024-04-11.

Sackmann, J. (2019). Jeff sackmann's tennis point-by-point data. Data file.

Sipko, M. and Knottenbelt, W. (2015). Machine learning for the prediction of professional tennis matches. MEng computing-final year project.

S.M. Ma, C.C. Liu, Y. T. and Ma, S. (2013). Winning matches in grand slam men's singles: An analysis of player performance-related variables from 1991 to 2008. *Journal of Sports Sciences*, 31(11):1147–1155.

Statista (2021). Market size of the online gambling industry worldwide from 2019 to 2023. Accessed: 2024-06-25.

Tennis Data (2024). Tennis data: Historical tennis results and odds. Accessed: 2024-08-13.

Tennis Planet (2020). Average career length of tennis player, nfl player, nba. Accessed: 2024-04-16.

Topspin, H. (2019). An introduction to tennis elo. Accessed: 2024-06-19.

Wilkens, S. (2020). Sports prediction and betting models in the machine learning age: The case of tennis. *Journal of Sports Analytics*. (Preprint), pp.1-19.