The SkillO Rating System

Nathan Dennis, John Breedis, Yiming Chen

Abstract

Predicting tennis match outcomes has become a popular application of rating systems, with the ELO system being widely used to assess players' skill. However, ELO cannot account for variability in a player's performance, which can lead to inaccurate predictions when players exhibit inconsistent behavior. To address this, we introduce SkillO, a new rating system incorporating Bayesian uncertainty into surface specific player ratings. This model combines ideas from the ELO and TrueSkill rating systems. SkillO adjusts tennis players ratings from match history between 2014-2022 while incorporating uncertainty in their rating through a variance parameter. We evaluate SkillO on 3 Grand Slam tennis tournaments in 2023, comparing its predicted champion probabilities to betting odds and the ELO system results. Results show SkillO outperforms ELO in certain scenarios, such as clay tournaments, highlighting the potential benefits of incorporating Bayesian uncertainty in these tournaments. Despite this, ELO remains generally more effective overall.

1 Introduction

The application of machine learning and mathematical models to predict sports outcomes has gained significant attention recently, with tennis match prediction presenting an interesting challenge [4]. The variability of individual player performance across different surfaces and conditions makes tennis an interesting domain for prediction models. One of the most popular methods to evaluate a players strength is through using rating systems, which quantify the skill level of players based on current and historical performance.

The ELO system is one of the most popular rating systems used to predict tennis matches [1]. Originally developed for chess, the ELO system adjusts player ratings after each match based on the skill level of the players involved. We utilize the ELO system in this project to predict tennis tournament champions, taking surfaces, tournament level, and match year as factors into the ELO calculation. While the ELO system is widely used and effective, it has some limitations. For example, the ELO rating system has no uncertainty or variability in a player's performance. Every player is given a single ELO rating, with no indication of how consistent they are as every outcome is based solely on the ratings of each player.

To address this key limitation, we introduce SkillO, a new rating system specifically designed to incorporate Bayesian uncertainty in predictions, a method that accounts for the variability in player performance. SkillO builds upon the ELO and TrueSkill rating system developed by Microsoft, originally used for video games [3, 6]. Unlike ELO, SkillO adjusts player ratings dynamically based on match outcomes and incorporates variability in their ratings through a variance rating parameter. The incorporation of variability in player ratings can help to improve match predictions by accounting for the inherent volatility in player's performances, especially when a player is playing more inconsistently where the ELO formula may not fully account for this behavior. We take influence from the TrueSkill

rating system, which incorporates uncertainty in player ratings through a variance parameter utilized in game score calculations between players and teams. We take influence from the dynamic K factor in the ELO calculation where we utilize the γ^2 parameter from TrueSkill to update player ratings based on surfaces, match year, and tournament level.

We evaluate the ELO and SkillO model on tennis tournaments in 2023. We utilize match data from the years 2014-2022 to calculate SkillO and ELO ratings for each player who has played a professional tennis match in these years. We make predictions on the 2023 Australian Open, Roland Garros, and Wimbledon to compare our models predicted champion probabilities to betting odds [8] using these rating systems and age of players. We hypothesize that by incorporating more variability in player ratings, SkillO will provide more accurate predictions than ELO for tennis tournaments dependent on parameter combinations.

2 Model Description

2.1 ELO Ratings Calculation

The first formula we used was similar to the ELO formula we developed in project 2. In this approach, we utilized the standard ELO formula and assigned players an ELO rating based on their match history. To begin, all players received an initial ELO rating of 1500 across all 3 tennis surfaces, Clay, Hard, and Grass. These ELO scores are iteratively adjusted based on past tennis match data between 2014-2022. We begin by setting the K factor to equal 20.

First, the K factor in the ELO calculation is adjusted based on the tournament level. The K factor is multipled by 4 for Grand Slam tournaments, multiplied by 2 for Masters tournaments, multiplied by $\frac{1}{2}$ for Davis Cup games, and kept the same for other matches, which imitates the well respected ATP ranking system. Furthermore, the K factor is adjusted based on year the match was played in. This logic is that older matches should be weighed less, more recent matches weighted more highly. The K factor is adjusted based on this using the formula: $K = Ke^{-0.3x}$ where x is the different in years, Current Year (In this case we train until 2022) - Year Match was played in.

Next, the probability of a player winning is calculated. We utilize the common expected game score formula used in ELO rating calculation given by:

$$P_{i,j} = \frac{1}{1 + 10^{(\frac{R_i - R_j}{800})}} \tag{1}$$

Where $P_{i,j}$ is the probability player i beats player j. Then R_i and R_j are the ELO ratings for players i and j on a surface. We also calculate $P_{j,i} = 1 - P_{i,j}$, the probability player j beats player i. The ELO scores for the given surface the match was played on for the winner and loser are adjusted as follows, under the assumption that player i defeats player j:

$$R_i' = R_i + K \times (1 - P_{i,j}) \tag{2}$$

$$R_j' = R_j + K \times (0 - P_{j,i}) \tag{3}$$

Furthermore, we understand that winning on one surface has an affect on the ratings for other surfaces. This is due to several reasons, such as a player may be playing better on one surface and this success may carry over to other surfaces. So, we adjust the players ELO ratings on the other 2 surfaces they didn't play on using a similar formula, just scaling the adjustment down slightly:

$$R_i' = R_i + K \times 0.8 \times (1 - P_{i,j}) \tag{4}$$

$$R_j' = R_j + K \times 0.8 \times (0 - P_{j,i}) \tag{5}$$

R_i' and R_j' are the new ratings on the 2 surfaces that were not played on. For example, if a match is played on Clay, the Clay surface ELO ratings for the players will adjust according to eqs. (2) and (3). The ELO ratings for the players Hard and Grass surface will adjust according to eqs. (4) and (5). These depend on who won the match to determine i and j. We decided to adjust with a 0.8 weight since through testing, weighing matches higher across all surfaces seemed too improve overall results.

2.2 SkillO Ratings Calculation

In this subsection, we present our new rating system, SkillO. This rating system is a combination of the common rating system, ELO, alongside a lesser known rating system, TrueSkill. The TrueSkill algorithm is based on Bayesian updating and helps estimate a player's skill level/rating while factoring in uncertainty about their skill. It incorporates 2 key parameters, the skill level of a player which follows a Gaussian distribution, and the uncertainty in the players rating. Each player, i, is assumed to have a real-valued skill denoted as $skill_i^t$ at time t, in our case the representing the number of matches a tennis player has played. The skill of a given player in a match is drawn from this distribution:

$$skill_i^{t_0} \sim N(m_0, \sigma_0^2) \tag{6}$$

Where t_0 is the time of the players first match and (m_0, σ_0^2) are the mean and variance parameters of a players skill. m_0 is the expected skill level of a new player and σ_0^2 the initial uncertainty about this skill. The common initial values are set to be $m_0 = 25$ and $\sigma_0^2 = \frac{25}{3} = 8.333$ [7], which will be iteratively adjusted based on past tennis match data between 2014-2022. Rather than draw from a Gaussian, in the SkillO approach we use the mean, μ, to represent the players skill level and σ^2 the uncertainty about this skill, with separate ratings on the 3 different tennis surfaces. In the TrueSkill system, after a given match the player's skill changes by a random amount drawn from a Gaussian:

$$skill_i^{t+1} \sim N(skill_i^t, \gamma^2) \tag{7}$$

Where $t+1$ represents a players skill after their next match and γ^2 is a tunable variance parameter controlling how much a player's skill can change after each match. Now, we instead use a similar approach with the ELO formula where the K factor is used and adjusted based on the year a match was played in and tournament level. We present a similar approach, where we begin with $\gamma^2 = 0.1$. Similar to the ELO K factor adjustment, this γ^2 learning rate is changed depending on the tournament level. It is multiplied by 4 for grand slams, 2 for masters, half for davis cup, and there is no rating change for other tournaments.

Furthermore, the γ^2 factor is adjusted based on year of the match. We decided to use a decay rate parameter of 0.7, where we adjust γ^2 using the formula: $\gamma^2 = \gamma^2 e^{-0.7 \cdot x}$ where x is the difference in years (current year - year match was played in). We decided on 0.7 since this would mean that 1 year in the past is weighted half of the current years ($e^{-0.7} \approx 0.5$), and the years further into the future are weighed much less, a larger decay than ELO.

To calculate the probability one player beat another, we used the following formula:

$$p_{winner} = \frac{1}{1 + e^{\left(\frac{\mu_{loser} - \mu_{winner}}{\sqrt{\sigma^2_{winner} + \sigma^2_{loser} + \beta^2}}\right)}} \tag{8}$$

μ_{loser} and μ_{winner} represent the current mean rating of the winning and losing player, σ^2_{loser} and σ^2_{winner} the current variance of the winning and losing player on a surface. We

utilized the sigmoid function rather than logistic for this calculation, as used in TrueSkill. The β parameter is used to control the effect of the variances in the calculation, limiting how highly variance is weighted. The parameter β was tuned in analysis to best reflect the context of tennis matches. We set the initial $\beta = 2$ after some experimentation since we observed a rapid decrease in calculated player variance with a high β like 4, but slow decrease with a low β, like 0.2. This causes small variance differences to give some players extremely low chances of winning a match with small β, so we set a rather conservative value for β to ensure the variance is not too highly influential in a match outcome. Equation (8) can also be thought of as player i beats player j, but to make the rest of the formulas easier to understand we opted to specify winner and loser. This was also an adjustment to the TrueSkill approach which did not end up using the β parameter in this calculation.

Then, the probability of the losing player winning is $p_{loser} = 1 - p_{winner}$. Next, we discuss how we update the mean and variance. Based on eq. (7), the TrueSkill formula updates player ratings based on the random γ^2, tunable parameter. In our approach, we specified that we developed a formula where we utilized γ^2 similar to the K factor in ELO calculation. To update the mean skill for the winning and losing player we utilized the following formula where we first calculated a γ^2 scaling factor to weigh each match, $\gamma^2_{scale} \sim |N(0, \gamma^2)|$, the absolute value of the Gaussian. This γ^2_{scale} is taken from a normal distribution with mean 0 and variance γ^2, where the γ^2_{scale} weight changes for every match in the dataset based on this distribution and value of γ^2. The means are updated using the following formulas:

$$\mu^{t+1}_{\text{winner}} = \mu^{t}_{\text{winner}} + \gamma^2_{\text{scale}} \cdot (1 - p_{winner}) \tag{9}$$

$$\mu^{t+1}_{\text{loser}} = \mu^{t}_{\text{loser}} + \gamma^2_{\text{scale}} \cdot (0 - p_{loser}) \tag{10}$$

Where μ^t_{winner} and μ^t_{loser} are the mean skill ratings on a surface of the winner and loser before the match. p_{winner} and p_{loser} represent the expected winning probabilities of the winner and the loser, calculated above. $t+1$ indicates the new mean rating after the match.

To update the variance, we use different techniques than discussed in the TrueSkill paper. Our approach updated the variance dependent on the expected winner/loser of each match.

$$\sigma^{2,t+1}_{\text{winner}} = \begin{cases} \sigma^{2,t}_{\text{winner}} \cdot (1 - \gamma^2_{\text{scale}} \cdot (1 - p_{\text{winner}})) & \text{if winner was expected to win} \\ \sigma^{2,t}_{\text{winner}} \cdot (1 + \gamma^2_{\text{scale}} \cdot p_{\text{winner}}) & \text{if winner won unexpectedly} \end{cases} \tag{11}$$

$$\sigma^{2,t+1}_{\text{loser}} = \begin{cases} \sigma^{2,t}_{\text{loser}} \cdot (1 - \gamma^2_{\text{scale}} \cdot p_{\text{loser}}) & \text{if loser was expected to lose} \\ \sigma^{2,t}_{\text{loser}} \cdot (1 + \gamma^2_{\text{scale}} \cdot (1 - p_{\text{loser}})) & \text{if loser lost unexpectedly} \end{cases} \tag{12}$$

We have $\sigma^{2,t}_{winner}, \sigma^{2,t}_{loser}$ as the variance rating on the surface played on for the winner and loser prior to the match. Our logic was that if the winner was expected to win, their variance should decrease since this was expected, however if the winner was not expected to win we are now more uncertain regarding their actual rating, hence their rating variance increases. If the loser was expected to lose their variance would decrease, but if they unexpectedly lost their variance would increase due being more uncertain in this players rating. We define an "unexpected win" if the player who won the match had under a 50% chance to win, but an expected win if the player who won the match had over a 50% chance to win based on p_{winner}. An "unexpected loss" is when the player who lost had over a 50% chance to win, but expected loss if they had under a 50% to win based on p_{loser}.

We used a similar approach to Equation (4) to change the rating and variance on other surfaces after a match for each player. The same γ^2_{scale} parameter in the mean calculation, Equations (9) and (10) and variance, Equations (11) and (12), is multiplied by 0.8 to adjust the ratings and variance on the other 2 surfaces using p_{winner} and p_{loser}.

Since this approach has a significant amount of randomness from γ^2_{scale}, we decided to run the simulation of calculating players SkillO ratings 30 times and averaged their means and variances across all 30 trials. Given using one simulation could harm analysis, this method is meant to use many simulation runs in order to accurately estimate a players rating.

2.3 Tournament Simulation

To simulate tournaments, both the ELO and SkillO approaches used similar methods with their respective rating systems. For ELO, the calculated winning probability for player i beating player j was given by eq. (1), the exact same equation used to calculate the expected game score when calculating ELO ratings for each player. Similarly in the SkillO approach, a slightly different formula, logistic, was used compared to eq. (8), given by:

$$P_{i,j} = \frac{1}{1 + 10^{\left(\frac{\mu_j - \mu_i}{\sqrt{\sigma_i^2 + \sigma_j^2 + \beta^2}}\right)}} \tag{13}$$

Which represents the probability that player i beats player j with the same β parameter when calculating players SkillO ratings. This is the key difference in calculating probabilities players beat one another, as the rest of the simulation is similar to the ELO formula. The corresponding probability player j beats player i is $P_{j,i} = 1 - P_{i,j}$ for both rating systems.

Using these probabilities, the probability both players win a given set is calculated. There are 5 sets in grand slam tennis matches, where the first player to win 3 sets wins the match. To compute the winning probability for each set in a given match for both players, we implement an age decay factor. The winning probability for a given set is given by:

$$\text{factor} = \begin{cases} 1 & \text{if age} \leq 25 \\ (e^{-\text{decay\_rate} \times (\text{age}-25)})^{Set-1} & \text{if age} > 25 \end{cases} \quad \text{where} \quad \text{decay\_rate} = \begin{cases} 0.015 & \text{if surface} = \text{clay} \\ 0.0075 & \text{otherwise} \end{cases}$$

We apply an age decay factor on the winning probabilities for players specific to the surface of the match, with a higher decay for Clay. We do this as in sports, older players experience faster skill decay due to increased fatigue compared to younger players. Through research, the Clay surface in tournaments such as the Roland Garros tend to be slower paced where the ball bounces higher than other surfaces and matches taking longer in general [5]. Because of this, we decided the decay rate for the Clay surface would be greater with longer matches. The probability a player wins a set from Equations (1) and (13) is multiplied by this factor, we see as the number of sets increases there is more decay in winning probability for older players. We normalize the probabilities between both players before calculating random uniform numbers to determine match winners in each set, first to 3 wins.

3 Analysis

We conduct analysis by observing the plots of predicted winning probabilities from 5000 simulations and table of errors of our ELO and SkillO model against betting odds. The ELO model stays constant from the described methodology. We experiment with different parameter combinations for the SkillO system based on each simulations results. For the first parameter combination: $\mu_0 = 25$, $\sigma_0^2 = 8.3333$ will be the initial mean and variance for players, $\gamma^2 = 0.1$ as the learning rate, and $\beta = 2$ the scaling factor for weighing the variance parameter when calculating winning probabilities. We calculate RMSE (Root Mean Squared Error), L_∞ (Largest absolute error), and L_1 (Total sum of absolute errors) to compare the SkillO and ELO results to the true value, betting odds. We plot the top 10 player's winning probabilities based on bettings odds alongside the ELO and SkillO champion probabilities.

3.1 Simulation 1

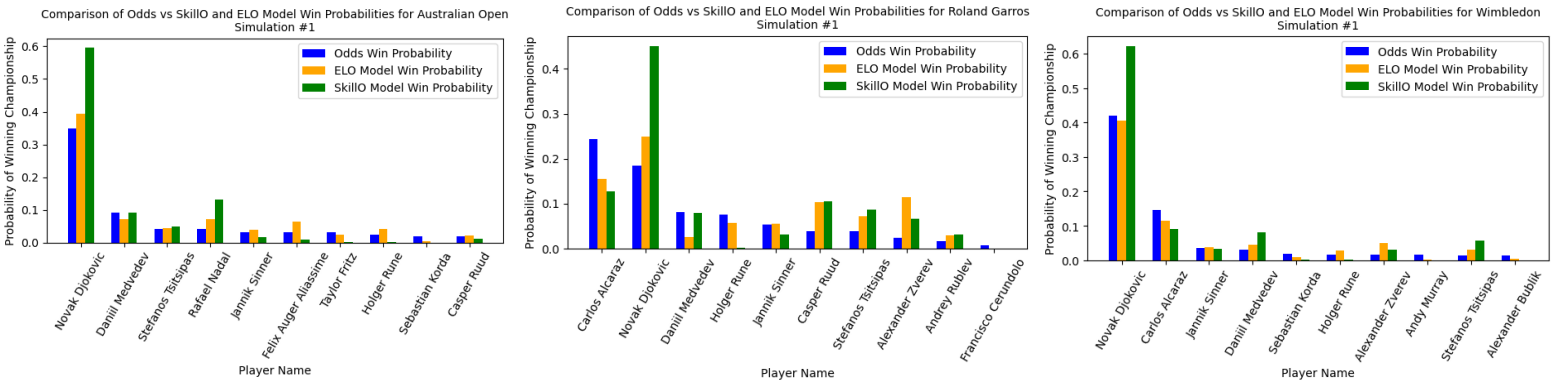

Figure 1: ELO vs SkillO Plots Across Tournaments, Simulation 1

	Australian Open			Roland Garros			Wimbledon		
Model	RMSE	L_∞	L_1	RMSE	L_∞	L_1	RMSE	L_∞	L_1
ELO	0.009428	0.074468	0.003624	0.016113	0.089464	0.005494	0.007039	0.032753	0.003529
SkillO	0.02529	0.246384	0.006205	0.029612	0.266258	0.007504	0.020966	0.201725	0.00574

Table 1: ELO vs SkillO Error Metrics across Tournaments, Simulation 1

We first observe in Figure 1 that the SkillO model is overall significantly more inaccurate than the ELO model comparing these results to the betting odds. It seems that consistently across all 3 major tournaments, the SkillO model heavily favors Novak Djokovic to win, much more than the odds and ELO. There are some players where the SkillO model predicts better than the ELO model, such as Daniil Medvedev. For Holger Rune the SkillO model performs much worse, basically giving him a 0 percent chance to win any tournament where the ELO rating system more properly rates him. We also noticed that Rafael Nadal has a significantly larger chance to win than the ELO model and betting odds, suggesting that our SkillO model heavily favors older players with this current parameter combination.

Looking at the results from Table 1, we again see that the SkillO model performed worse in terms of RMSE, L_∞ and L_1 score across all 3 tournaments compare to ELO. The largest difference was in the Australian Open, as the SkillO rating system was significantly worse than the ELO system across all metrics by a wide margin. It seems like SkillO performed best in Wimbledon, but compared to ELO it was still significantly worse. SkillO performed the worst on Clay, similar to ELO, with one prediction that had nearly a 27% difference from the betting odds, which we can see from Figure 1 was Novak Djokovic.

From this analysis, we decided to set a larger decay factor when running the SkillO rating calculation from past years. Based on Figure 1, it was clear that the SkillO rating system favors older players, as seen with Novak Djokovic and Rafael Nadals high winning probabilities compared to both the ELO results and betting odds. We also noticed that the variance for both Djokovic and Nadal was very low from the SkillO rating dataframe, indicating that potentially the β parameter should be tuned further when calculating SkillO ratings as this parameter directly affects winning probabilities.

3.2 Simulation 2

We now experiment with new parameters for the SkillO rating system. We still use the same initial mean and variance, $\mu_0 = 25$, $\sigma_0^2 = 8.3333$ as the starting mean and variance for players, $\gamma^2 = 0.1$ as the learning rate parameter. Now, we change the year decay rate

parameter to be 1.1, which indicates a faster decay for matches played further in the past, where matches played 1 year ago are decayed by $e^{-1.1} \approx \frac{1}{3}$ when calculating SkillO ratings. We also adjust β when simulating SkillO ratings and tournaments, setting $\beta = 1$. As stated in the model description, smaller β led to a slower decrease in variance. We hope this change will make the variances of players more even and have less of a discrepancy with older players such as Djokovic or Nadal. We now view these figures and corresponding table of errors.

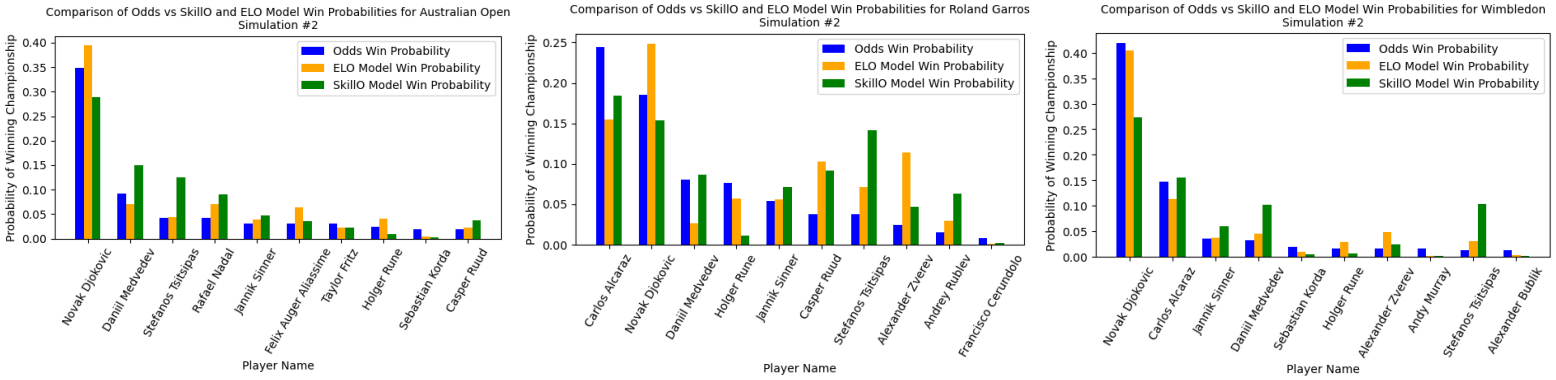

Figure 2: ELO vs SkillO Plots Across Tournaments, Simulation 2

	Australian Open			Roland Garros			Wimbledon		
Model	RMSE	L_∞	L_1	RMSE	L_∞	L_1	RMSE	L_∞	L_1
ELO	0.009428	0.074468	0.003624	0.016113	0.089464	0.005494	0.007039	0.032753	0.003529
SkillO	0.013814	0.082791	0.005005	0.015135	0.103124	0.005067	0.018349	0.144475	0.005452

Table 2: ELO vs SkillO Error Metrics across Tournaments, Simulation 2

Based on the plots from Figure 2, we can see vastly improved predictions compared to Figure 1. The predicted probability of Novak Djokovic winning is much lower, slightly closer to the betting odds for the Roland Garros compared to ELO. The SkillO system is more accurate than the ELO formula for several players across different tournaments, such as Novak Djokovic in the Roland Garros and Carlos Alcaraz in the Wimbledon. However, players such as Stefanos Tsitsipas are highly overrated across all tournaments by the SkillO model, whereas the ELO model is generally much more accurate. This may be due to his recent success, as SkillO could be weighing his success in the most recent year too highly.

In Table 2 we see a similar trend with the plots, the accuracy has significantly improved for SkillO and is much more comparable with the ELO formula. We can see the ELO formula still outperforms the SkillO model in terms of RMSE across all surfaces except for Roland Garros on clay, which prompts the question if SkillO may be better than ELO on clay. The L_1 scores for the SkillO model are nearly the same as the ELO formula, indicating the average absolute differences between the models predictions are not too heavily far apart. Furthermore, the L_∞ scores are also very similar except for Wimbledon, where SkillO is highly inaccurate by almost 15% for one player, which was Stefanos Tsitipapas from Figure 2.

Based on this analysis, the next move we decided to make was adjusting the γ^2 parameter, which is sort of like the K factor in the ELO rating system. We experiment what happens when increasing this γ^2 parameter, as currently we have it set to be $\gamma^2 = 0.1$. We experiment with a larger value of $\gamma^2 = 0.15$. We are interested if a slightly larger scaling factor can help solve the discrepancies in giving players like Stefanos Tsitipapas higher ratings than they should have. We do note that with larger γ^2 values like 0.5, the results are heavily inaccurate as it gives Novak Djokovic very high (> 0.7) chance to win each tournament. Furthermore, with smaller values like $\gamma^2 = 0.05$, the SkillO model underrates the top players significantly and gives more uniform probabilities for players to win each tournaments.

3.3 Simulation 3

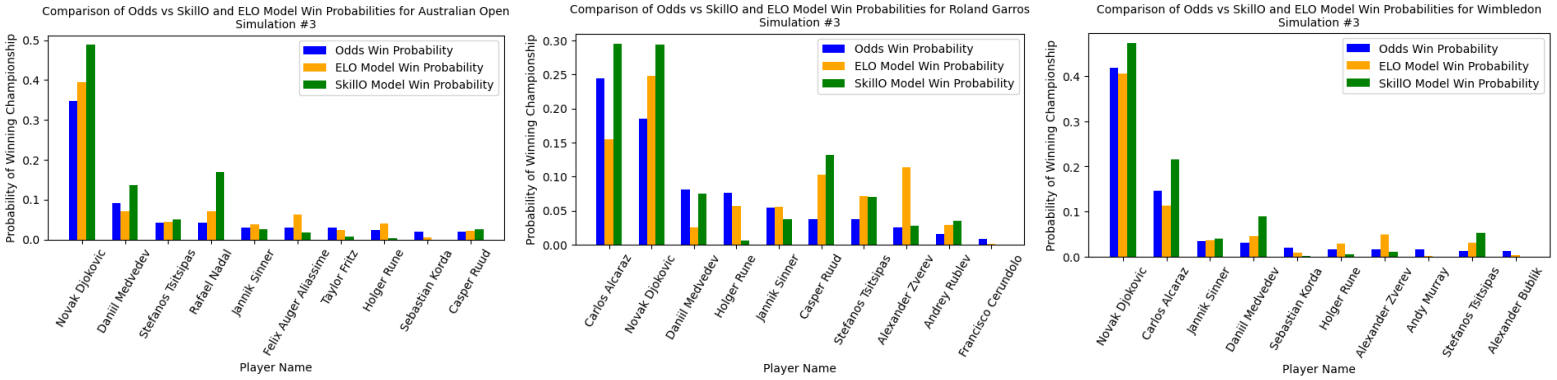

Figure 3: ELO vs SkillO Plots Across Tournaments, Simulation 3

	Australian Open			Roland Garros			Wimbledon		
Model	RMSE	L_∞	L_1	RMSE	L_∞	L_1	RMSE	L_∞	L_1
ELO	0.009428	0.074468	0.003624	0.016113	0.089464	0.005494	0.007039	0.032753	0.003529
SkillO	0.018997	0.139584	0.005719	0.016383	0.109258	0.005241	0.011484	0.068196	0.004578

Table 3: ELO vs SkillO Error Metrics across Tournaments, simulation 3

Based on the plots in Figure 3, we can see that the SkillO rating system seems to not follow the ELO rating as much. In the Roland Garros for example, the SkillO system is able to more accurately predict Carlos Alcaraz's high chance of winning compared to the ELO system. Still, the SkillO system consistently overrates players chances of winning more than the ELO system does, such as Novak Djokovic across all 3 tournaments. The SkillO system still seems to be more top heavy than the ELO system, as it gives the best players significantly higher odds to win compared to the lower level players which can especially be seen in the Wimbledon and Australian Open plots. Compared to the ELO rating system, SkillO rarely predicts lower ranked players to win the Australian Open, assigning very high probabilities to the top players. For the Roland Garros, the SkillO formula seems to perform well overall but still highly overrates players compared to ELO, such as Djokovic or Ruud.

As for Table 3, we actually see that the SkillO model performs near the same as the ELO model in the Roland Garros. The errors for the Australian Open are smaller and better for the ELO model across all error metrics, but the SkillO system could be improved further if it didn't overrate Nadal so much from Figure 3. The Wimbledon errors are more similar to the ELO errors, though ELO still outperforms SkillO across all metrics. The most noticable similarities are for the Roland Garros, where the SkillO system performs very well and is comparable to the ELO rating system. While the ELO rating system performs better than ELO across the surfaces, the similarities in the Roland Garros prompt further analysis into whether the Clay surface could benefit with a rating system that incorporates uncertainty.

3.4 Overall SkillO Parameter Analysis

Now we compare the 3 SkillO modelling results and observe the differences between these approaches and parameter combinations.

We note that all 3 simulations used the same starting parameters, $\mu_0 = 25$, $\sigma_0^2 = 8.3333$. The first simulation uses $\beta = 2$ and $\gamma^2 = 0.1$ with a year decay rate of 0.7. The second simulation uses $\gamma^2 = 0.1$, but changes the year decay rate to be 1.1 and $\beta = 1$. The third simulation changes only $\gamma^2 = 0.15$ from the second simulation, keeping $\beta = 1$ and the same year decay factor.

	Australian Open			Roland Garros			Wimbledon		
Model	RMSE	L_∞	L_1	RMSE	L_∞	L_1	RMSE	L_∞	L_1
SkillO, Simulation 1	0.02529	0.246384	0.006205	0.029612	0.266258	0.007504	0.020966	0.201725	0.00574
SkillO, Simulation 2	0.013814	0.082791	0.005005	0.015135	0.103124	0.005067	0.018349	0.144475	0.005452
SkillO, Simulation 3	0.018997	0.139584	0.005719	0.016383	0.109258	0.005241	0.011484	0.068196	0.004578

Table 4: SkillO Error Metrics Across all Simulations

We can see from Table 4 that the SkillO rating system performed best in the second simulation across all 3 metrics for the Australian Open and Roland Garros, but the final simulation performed the best in the Wimbledon. It seems that the first model was by far the most inaccurate. The second and third simulation with $\beta = 1$ did much better than the first simulation with $\beta = 2$, indicating that a smaller β value could be beneficial, but a larger β value may heavily deflate player variances. Furthermore the higher decay factor for year seemed to improve results. From Figure 1 we saw that across all 3 tournaments Novak Djokovic was heavily favored to win compared to the ELO system and betting odds, but increasing the year decay factor for an older player such as Novak decreased his probability of winning and made the overall results more comparable. It seems like changing the γ^2 parameter, increasing it specifically in the third simulation, actually slightly improved results for the Wimbledon, indicating that each tournament may have specific parameter combinations that could be tuned specifically for that tournament.

The L_1 and L_∞ scores improved in the second and third model compared to the first, indicating the predicted probabilities were much closer with these models compared to betting odds. The second and third simulation had comparable L_1 and L_∞ scores outside of Wimbledon, implying again that different tournaments may require tuning parameters such as γ^2 to achieve the best results. Overall, it seems like changing the β parameter and decay rate heavily influenced results, as the first simulation might have had inadequate parameters.

4 Discussion

Through this analysis we were able to compare our new SkillO rating system to both the ELO rating system and betting odds. We were not surprised that the ELO model usually performed better than SkillO, as the ELO model is a very well respected rating system that many people use. It is noteworthy that SkillO showed slightly better performance at predicting the Roland Garros in the second simulation, suggesting that incorporating Bayesian uncertainty like the SkillO model uses can improve prediction accuracy on the clay surface. As for the other tournaments, the ELO model was relatively accurate and could indicate these tournaments are easier to predict. It may not be necessary to utilize a rating system that incorporates Bayesian uncertainty for these tournaments, but future work could investigate optimizing specific parameter combinations for each tournament rather than a single parameter combination for all 3 to yield more accurate results for SkillO.

In terms of the parameters of the model, we were surprised that changing the year decay factor and β parameter significantly improved results. A small year decay factor may cause the SkillO model to weigh past matches too highly when calculating SkillO ratings, where older matches can give a false sense of how strong a player is performing in the present day. The β parameter may be the most influential in our model, as it controls how impactful the difference in variance between players is when simulating matches and calculating winning probabilities. A smaller β allows the players variances to contribute more to the overall prediction, where we found that a slightly smaller β of 1 improved results. This aligns will the SkillO model's core innovation, incorporating a variance parameter for each player to reflect the models uncertainty about their skill set. We were not surprised that increasing the γ^2

parameter overrated dominant players like Djokovic. This parameter directly influences the weight matches have on rating calculations and could overrate players who win consistently.

There are also several limitations with this new SkillO model. First, there is some randomness in the model, especially with how the γ^2_{scale} parameter is calculated, as it is essentially taken from a normal distribution and can weigh matches much differently. This is a part we utilized from the TrueSkill system, but realize it may incorporate too much unnecessary randomness into our model and future work could investigate changing this feature as results depend on this randomness. Another limitation is the scope of the project. We only evaluated the SkillO system on tennis tournaments, comparing it to both the ELO rating system and betting odds. Future work could compare the SkillO rating system to other popular rating systems, such as the Glicko method which is also widely used to model tennis tournaments and incorporates Bayesian uncertainty in the model [9, 2]. Future work could also use this rating system to model other tournaments, such as the NBA playoffs.

5 Conclusion

This analysis explored the limitation of the ELO rating system and introduced SkillO as an alternative which incorporates bayesian uncertainty to model player performance and variability. While the ELO model generally outperformed SkillO in most tournaments, SkillO still shows promise in some areas such as the Roland Garros. This prompts future analysis into which models are better for specific tournaments, utilizing different modeling approaches for separate tournaments. These findings also indicate fine-tuning parameters in the SkillO model, such as the year decay and β parameters, significantly impact the accuracy of predictions. SkillO relies heavily on randomness, especially with the γ^2 parameter, where future work could focus on refining this randomness and optimizing parameters for specific tournaments. Expanding the comparison to other rating systems and applying the model to a broader range of sports could provide valuable insights to improve SkillO's predictive power.

6 Risk

Through this project, the main risk we took was trying to develop our own rating system, SkillO, combining ideas from both ELO and TrueSkill. The TrueSkill system, which was originally designed for ranking players in Xbox games, represented an unconventional source of inspiration for a sports based rating system. While we found some of TrueSkill's ideas interesting, we didn't simply apply it to tennis prediction. Instead, we took influence from this approach, combining it with the ELO system and our own ideas to make a completely new rating system. This was risky, as taking influence from a rating system designed for Xbox players may not guarantee it would produce meaningful results predicting sports outcomes.

The riskiest and most challenging part of this project was designing the rating system, particularly deciding what factors we wanted to adopt from ELO or TrueSkill. We discussed as a group every step to design the system, testing various ideas from the ELO and TrueSkill system before finalizing on the final approach we developed in the Model Description section. We learned how difficult it truly is to develop your own rating system, especially when taking influence from a system that is not traditionally used for sports and one that doesn't have significant research conducted on it. While we were hoping the SkillO system performed better, we were quite pleased overall with its improvement over the repeated simulations and parameter combinations through our analysis. We learned through these risks that designing a rating system must take careful consideration of parameters and modeling approaches, as the parameters of the model are an important feature in any rating system.

7 Attribution of Effort

We worked together to develop the SkillO rating system. We all discussed what ideas we could take from the ELO and TrueSkill approach, as well as ideas we had to develop this new rating system. Basically, Nathan, John, and Yiming all worked together to develop the model. After we discussed and finalized our approach, we discussed some next steps we could take in analyzing the model, where we decided to compare our approach to our ELO model developed in past projects. In terms of work, we all worked together on the design but Nathan worked on writing the github code and commiting the code to our repository as well as the test cases for the code.

We note that Michael stated it was ok if the attribution of effort section went over the 10 page limit.

References

[1] Rory Bunker et al. *A Comparative Evaluation of Elo Ratings- and Machine Learning-based Methods for Tennis Match Result Prediction*. Sept. 2023.

[2] Yansong Dong. "Intelligent Analysis and Predictive Modeling of Tennis Match Data". In: *Applied Mathematics and Nonlinear Sciences* 9 (July 2024). DOI: 10.2478/amns-2024-1593.

[3] Ralf Herbrich, Tom Minka, and Thore Graepel. "TrueSkill™: A Bayesian Skill Rating System". In: *Advances in Neural Information Processing Systems 20*. Accessed: 2024-12-06. MIT Press, Jan. 2007. URL: https://papers.nips.cc/paper/2007/hash/61d0d6c5e953cf68abfbe90d6de0f50a-Abstract.html.

[4] Yilin Lei, Ao Lin, and Jianuo Cao. "Rhythms of Victory: Predicting Professional Tennis Matches Using Machine Learning". In: *IEEE Access* PP (Jan. 2024), pp. 1–1. DOI: 10.1109/ACCESS.2024.3444031.

[5] Caroline Martin and Jacques Prioux. "Tennis Playing Surfaces: Effects on Performance and Injuries". In: *Journal of Medical Sciences in Tennis* 20.3 (2015), pp. 1–10.

[6] Tom Minka, Ryan Cleven, and Yordan Zaykov. "TrueSkill 2: An improved Bayesian skill rating system". In: (Mar. 2018). Accessed: 2024-12-06. URL: https://www.microsoft.com/en-us/research/publication/trueskill-2-an-improved-bayesian-skill-rating-system/.

[7] Microsoft Research. *TrueSkill Ranking System*. Accessed: 2024-12-12. 2006. URL: https://www.microsoft.com/en-us/research/project/trueskill-ranking-system/.

[8] Sports Odds History. *Tennis Odds*. Accessed: 2024-12-14. 2024. URL: https://www.sportsoddshistory.com/tennis-odds/.

[9] Jack Yue et al. "A study of forecasting tennis matches via the Glicko model". In: *PLOS ONE* 17 (Apr. 2022), e0266838. DOI: 10.1371/journal.pone.0266838.