How well do Elo-based ratings predict professional tennis matches?

Abstract

This paper examines the performance of five different measures for forecasting men's and women's professional tennis matches. We use data derived from every match played at the 2018 and 2019 Wimbledon tennis championships, the 2019 French Open, the 2019 US Open, and the 2020 Australian Open. We look at the betting odds, the official tennis rankings, the standard Elo ratings, surface-specific Elo ratings, and weighted composites of these ratings, including and excluding the betting odds. The performance indicators used are prediction accuracy, calibration, model discrimination, Brier score and expected return. We find that the betting odds perform relatively well across these tournaments, while standard Elo (especially for women's tennis) and surface-adjusted Elo (especially for men's tennis) also perform well on a range of indicators. For all but the hard-court surfaces, a forecasting model which incorporates the betting odds tends also to perform well on some indicators. We find that the official ranking system proved to be a relatively poor measure of likely performance compared to betting odds and Elo related methods. Our results add weight to the case for a wider use of Elo-based approaches within sports forecasting, as well as arguably within the player rankings methodologies.

Key words: Forecasting, Elo, betting, tennis, calibration, expected return, Brier Score, prediction accuracy, model discrimination.

1. Introduction

The purpose of this paper is to examine the performance of different forecasting methodologies for both men's and women's professional tennis matches. The measures we use are the betting odds, the official men's tennis and women's tennis rankings, the standard Elo ratings, the surface-specific Elo ratings, and a composite of some of the above. The Elo rating system is a method of ranking players based on their past matches, weighted by the ratings of the players they competed against. The performance indicators we use are prediction accuracy, calibration, model discrimination, Brier score and expected return.

We focus on both men's and women's singles matches for the 2018 and 2019 Wimbledon tennis championships, the 2019 French Open, the 2019 US Open, and the 2020 Australian Open, employing data derived from every match played at these 'Grand Slam' tournaments.

Both the men's and women's singles in each tournament consist of 128 players, with direct entries based on the official Association of Tennis Professionals (ATP) rankings and the official Women's Tennis Association (WTA) rankings. Additional players of each gender are then chosen as 'wild card' entries, based on a player's previous performances during the season or by being a competitor of public interest to increase publicity for the event. The remaining spots are filled by the winners of qualifying matches held in the week prior to the main competition. The top-ranked 32 players of each gender are 'seeded' so that the best-ranked players do not play each other too early in the tournament. The rest of the players are then randomly assigned their matches, both against themselves and the top-ranked players.

The players compete in a “single elimination tournament modus (knockout system)” (Leitner et al., 2009, p. 278).

2. Literature

Stekler et al. (2010) provide a review of sports forecasts – see also Vaughan Williams and Stekler (2010) – noting that if we view betting odds as forecasts, then standard tests of forecast efficiency are also tests of information efficiency. Such studies have been common over the years – seminal papers include Snyder (1978), Asch et al. (1984) for horse race betting and Pope and Peel (1989) for football betting. Indeed, many forecasting methods are evaluated according to whether they would achieve positive betting returns – seminal papers include Vergin and Scriabin (1978) for American football, Bolton and Chapman (1986) for horse racing, while much more recently Angelini and De Angelis (2019) assess betting market efficiency for eleven European football leagues.

Among statistical forecasting models, a common approach is to rank participants based on historical performance. Many sports run official ranking systems, and in addition Elo (1978) proposed a rating system for chess that has been used in a range of sports. Hvattum and Arntzen (2010) test Elo ratings against bookmakers and econometric models as a forecasting tool for English Premier League matches, finding that bookmakers outperform Elo ratings, but that Elo ratings are superior to econometric models, while Leitner et al. (2010) use Elo ratings among other methods when attempting to forecast outcomes from the 2008 European Championships football tournament. Ryall and Bedford (2010) create an Elo-based model for Australian Rules football, and Carbone et al. (2016) do so for rugby league.

Kovalchik (2016) evaluates an Elo-based prediction system created by the website FiveThirtyEight.com (Silver and Fischer-Baum, 2015; Morris et al., 2016) and finds that this comes closest among a range of forecasting methodologies to beating bookmaker prices in tennis. Kovalchik and Reid (2019) extend this method for in-play tennis betting. Our study complements the work of Kovalchik (2016) - see also Kovalchik and Reid (2019) - in developing our own adjusted Elo ratings designed to improve forecasting performance of tennis matches, in our case for both men’s and women’s tennis across the four Grand Slam tennis tournaments. We develop explicit surface-specific Elo ratings, as well as using standard Elo ratings.

3. Methodology

The metrics we use are the betting odds, the official men’s (ATP) and women’s (WTA) tennis rankings, and Elo related ratings.

3.1 Betting odds

To find the best odds available for the analysis, the odds comparison site, Oddschecker (2018, 2019, 2020) was used as it collates all the data from a range of betting operators to highlight the best available odds. The odds were deflated by the over-round (the excess of the sum of the implied probabilities in the odds over 1) to give the implied probabilities for each player in a match. Regarding the fractional odds, the method by which these probabilities were calculated is given in Equation (1), which follows Graham and Stott (2008). See also Clarke et al. (2017).

$$p = \frac{denominator}{denominator + numerator} * 100 \qquad (1)$$

3.2 Association of Tennis Professionals (ATP)/Women’s Tennis Association (WTA) rankings

The ATP and WTA official world rankings, for men and women’s tennis respectively, are used within professional tennis to determine tournament eligibility. They both follow a 52-week cumulative rolling points system, with the results from the four Grand Slam tournaments having the highest points weighting. The weighting of the points increases with the prestige of the tournament, as well as the round of the tournament reached. The points accrued from 19 ATP and 16 WTA tournaments out of all those played (weakest tournament scores drop out) are totalled to create the overall rankings of the players (Dingle et al. 2012).

3.3 Elo

The Elo rating system, originally developed by Arpad Elo (Elo, 1978) as a method of ranking chess players, takes the relative skill level of players based on their past performances to establish a prediction for a head-to-head outcome, and then updates the ratings after each match result.

The method works by allocating more points to a player when defeating a stronger opponent and deducting points when losing to a weaker opponent (Hvattum and Arntzen, 2010).

As a general rule, a 100-point difference is the equivalent of a 64% chance of winning, a 200-point difference equivalent to 75%, and 300-point difference to an 85% chance (Walkofmind, 2019) - see Equation (2).

$$p_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{2}$$

R_A and R_B are the ratings for player A and B. The Elo rating differences were converted to win probabilities (p_A) for each player in a match. The use of 400 is widely used in chess organizations. Tennis Abstract (2020) also calibrate this number to be 400 to reflect that a 100-point difference in Elo ratings implies that the favorite has a 64% chance of winning.

With this win probability, Player A's new rating score ($R_A^{'}$) can be updated using Equation (3).

$$R_A^{'} = R_A + K(S_A - p_A) \tag{3}$$

where S_A is the actual score for Player A and K is a factor to determine the amount by which the Elo rating should be updated after each match. If the K-factor is high, the new rating responds with high sensitivity to the performance. If the K-factor is low, the sensitivity of the adjustment is small. In practice, there are three types of approaches to set this value. Firstly, and originally, the K-factor was set to be 10 for players with ratings above 2400. Sonas (2002) argued, however, that $K=10$ is an inaccurate reflection of a player's actual level. He proposed a K of 24 based on empirical observations derived from actual matches. Secondly, the K-factor was set to be different across different levels. The International Chess Federation (FIDE), for example, uses K = 40, K = 20, and K = 10 based on player ratings, the number of games completed and the player's age. Lastly, the K-factor is set according to a continuous function rather than a constant, such as in the United States Chess Federation (USCF) system.

Standard and surface-specific Elo ratings, which are the official ratings, were used within the methodology:

1. Standard Elo for ATP and for WTA.

2. Surface-specific Elo. Wimbledon is played on a grass court, so a surface-specific Elo only accounts for games played by the competitors on a grass surface. The French Open is played on a clay-court surface and the US Open and Australian Open on hard-court surfaces.

3.4 Adjusted Elo ratings

We find that the official ranking proved to be a relatively poor measure of likely performance, highlighting a possible case for a change in the method by which the official rankings are calculated (see also Reid et al., 2010). An adjusted/combined Elo is proposed in this paper to improve the forecasting performance of tennis matches. This weights both standard and surface-specific Elo. As Wimbledon, for example, is played on a grass-court surface, the grass surface ratings are chosen to reflect the player's abilities within this match scenario. We construct an adjusted Elo rating to reflect both Elo and surface ratings, which is shown in Equation (4).

$$Adjusted\ Elo\ 2 = (1 - \lambda) * StandardElo + \lambda * SurfaceElo \tag{4}$$

The simplest adjustment is to weight each type of Elo equally, so taking the midpoint of the standard Elo and surface-specific Elo for each player (Adjusted Elo ratings 1). However, the equal weight of Elo and surface-specific Elo may not be optimal. Considering this, we set λ to be varying between 0 and 1. For each λ, we calculate the prediction accuracy, calibration, model discrimination, Brier score and expected return. We choose the maximum value (best performance) of these measures. The corresponding λ is the optimal weight on surface-specific Elo. Instead of placing equal weights on Elo and surface Elo, we have calculated the adjusted Elo ratings (Adjusted Elo ratings 2), which uses the optimal weights. As the actual outcome and existing Elo ratings may not be linearly related, we borrowed the idea of Indirect Inference estimation (see Smith, 1993) to estimate these weights rather than applying OLS.
As the forecasting performance of betting odds is another important indicator, we extend the current literature by constructing another rating in the Equation (5) incorporating the betting odds.

$$Adjusted\ Elo\ 3 = (1 - \lambda_1 - \lambda_2) * StandardElo + \lambda_1 * SurfaceElo + \lambda_2 * Betting\ Odds \tag{5}$$

We set λ_1 and λ_2 to be varying between 0 and 1 but the sum of them cannot exceed 1. For each combination, we calculate the forecasting measures. The ones that maximize the forecasting performance are the optimal values for these weights.

The idea of developing a weighting-based or rule-based combination of methods to improve forecasting accuracy in sport has been previously explored by, for example, Spann and Skiera (2009) but not applied in this way.

4. Model performance

To test the performance of the models, five measures were used: prediction accuracy, calibration, model discrimination, Brier score and expected return. When looking at the predictive power of a model, although accuracy may be viewed as the most desirable characteristic, the sensitivity to bias within the model is also important (Irons et al. 2014), hence the choice of these different measures.

Prediction accuracy is a measure of the number of correctly predicted matches that the player with the higher probability won. It is calculated by finding the number of matches that were correctly predicted divided by the total number of predictions and is expressed as a percentage.

$$Prediction\ accuracy = \frac{total\ number\ of\ correctly\ predicted\ matches}{total\ number\ of\ predictions} * 100 \tag{6}$$

Calibration can be defined as how well the forecasted probabilities correspond to the actual outcomes (Tetlock and Gardner, 2015). In this paper, a calibration ratio is used, calculated as the sum of the probabilities of the higher-ranked player winning divided by the number of matches the higher-ranked player won.

$$Calibration = \frac{\text{sum of the probabilities of the higher ranked player wins}}{\text{total number of matches the higher ranked player won}} * 100 \tag{7}$$

The closer the ratio is to 1, the better calibrated and less biased the model is. If the model puts more weighting on the higher-ranked players to win, the calibration will be more than 1, with a model underestimating the higher-ranked players having a ratio less than 1.

Model discrimination is calculated as the mean probability of matches the higher-ranked player won minus the mean probability of when they lost (upsets).

$$\begin{aligned} & Model\ discrimination \\ & \qquad = mean\ prediction\ for\ matches\ higher\ ranked\ player\ won \\ & \qquad - mean\ prediction\ for\ matches\ they\ lost \end{aligned} \tag{8}$$

This is equivalent to the integrated discrimination improvement (IDI) measurement used by Pencina, D'Agostino and Vasan (2008). Higher values of the IDI and model discrimination reflect a higher discriminatory power, indicating that the probabilities are more certain for wins than upsets within the matches.

The Brier score is another way to measure the prediction accuracy, which is between 0 and 1. It is an average sum of the squared difference between a predicted probability and actual outcome of all matches. The higher the Brier score is, the worse the prediction is.

$$Brier\ score = \frac{1}{N}\sum_{i=1}^{N}(probablity\ of\ forecast - outcome)^2 \tag{9}$$

N is the number of matches recorded. For each match, the probability that a particular player wins is calculated using the betting odds comparison site, Oddschecker. If the player wins, the outcome is 1; if the player loses, the outcome is 0. The difference between forecasting probability and the actual outcome can then be calculated for each match. We take the average of the squared difference to measure this forecasting accuracy.

Finally, we calculate the expected return to bets placed on players whose implied win probability in a match based on Elo ratings exceeds that implied in the betting odds. There is an extensive literature that suggests that sports betting markets (including tennis betting markets) are indeed efficient or close to efficient (e.g. Reade et al., 2020; Easton and Uylangco, 2010; Vaughan Williams, 2005), and we might expect the weight of informed money to drive the odds to closely reflect the true implied probabilities of winning. As such, we consider that expected return is a useful additional measure of model performance. To calculate expected return, we place a notional unit stake in all matches where the implied probability that a player will win based on the Elo ratings exceeds the probability implied in the odds. In other words, the same amount of capital is staked on every player whose implied probability of winning based on their Elo rating is greater than the implied probability in the betting odds. The idea is that these players are more likely to win than the betting odds imply, and so we are obtaining good value. If the implied win probability of the player based on the Elo ratings is smaller than the implied probability in the betting odds, no bet is placed. The total number of matches used in Equation (10) is, therefore, smaller than all the matches observed. The implied probability in the betting odds can be calculated by Equation (1), while the probability implied in the Elo ratings can be determined by Equation (2). Suppose the fractional odds of Player A is 2/1. In this case, the net profit is twice the unit stake if the player wins, but the net profit is minus the unit stake if the player loses. A higher expected return indicates better forecasting performance.

$$\begin{aligned} & Expected\ return \\ & = \frac{total\ profit\ when\ implied\ Elo\ probability\ is\ greater\ than\ implied\ odds\ probability.}{total\ capital\ when\ implied\ Elo\ probability\ is\ greater\ than\ implied\ odds\ probability.} \end{aligned} \tag{10}$$

5. Data

Table 1 summarizes the source and sample size of the data including men's Association of Tennis Professionals (ATP) rankings and women's World Tennis Association (WTA) rankings, betting odds, and Elo ratings. Data was collected for the ATP and WTA rankings, for the Elo ratings at the start of each tournament and for the betting odds before the beginning of play on each day of the tournaments. The ATP and WTA rankings were collected from the official websites, atpworldtour.com and wtatennis.com, respectively. The Elo and surface-specific Elo ratings were collected from Tennis Abstract (2018 -2020a, 2018-2020b). To find the best betting odds available, the betting comparison website, Oddschecker (2018, 2019, 2020) was used as it collates all the data from a wide range of betting operators to give the most competitive odds. Match results and information were obtained from Flashscore (2018, 2019, 2020).

Table 1: Summary of the data set

Data set	Source
ATP Rankings	ATP World Tour
WTA Rankings	WTA Tennis
ATP betting odds	Oddschecker
WTA betting odds	Oddschecker
ATP Elo ratings	Tennis Abstract
WTA Elo ratings	Tennis Abstract

Table 2. Summary statistics men's tennis

Variable	Obs	Mean	Std Dev	Min	Max	Tournament
ATP	139	77.5	52.3	1.0	256	Wimbledon
Elo	169	1738.5	131.9	1516.6	2222.3	2018
Elo Grass	169	1531.0	129.7	1209.1	1940.9	
ATP	252	60.7	52.6	1.0	286	Wimbledon
Elo	249	1863.3	156.9	1471.1	2188	2019
Elo Grass	247	1551.6	162.8	1187.0	1964.7	
ATP	253	69.0	65.6	1.0	260	US
Elo	248	1807.5	156.4	1461.8	2200.7	2019
Elo Hard	248	1694.8	172.0	1161.7	2079.9	
ATP	253	65.2	64.6	1.0	255	Australian
Elo	250	1811.2	175.0	1423.0	2222.5	2020
Elo Hard	250	1705.2	179.5	1228.2	2110.4	
ATP	253	62.9	58.3	1.0	273	French
Elo	247	1807.2	165.6	1475.4	2190	2019
Elo Clay	247	1697.1	185.8	1231.6	2127.6	

Table 3. Summary statistics women's tennis

Variable	Obs	Mean	Std Dev	Min	Max	Tournament
WTA	155	90.8	66.9	1	297	Wimbledon
Elo	173	1720.5	135.5	1425.4	2129.4	2018
Elo Grass	173	1514.1	119.9	1239.2	1797.5	
WTA	250	59.8	50.9	1	298	Wimbledon
Elo	246	1811.2	146.7	1412.3	2178.7	2019
Elo Grass	246	1527.1	137.4	1218.1	1842	
WTA	246	61.1	55.1	1	280	US
Elo	238	1822.0	139.7	1510	2126.6	2019
Elo Hard	238	1729.2	146.7	1416.1	2032	

WTA	251	59.3	57.6	1	226	Australian
Elo	247	1813.2	139.2	1426.9	2123.7	2020
Elo Hard	247	1719.6	145.6	1333.5	2031	
WTA	252	67.7	73.7	1	289	French
Elo	244	1808.0	147.4	1456.3	2116.8	2019
Elo Clay	244	1628.0	153.4	1107.8	1994.6	

6. Results analysis

6.1 Wimbledon 2018 and 2019

Figure 1 shows the forecasting performance over the two Wimbledon tennis tournaments combined using different rating methods. For men's tennis, we find that the betting odds outperform the other metrics in terms of prediction accuracy, calibration, model discrimination and Brier score. A simple weighted average of overall and surface-specific Elo performs best in terms of expected return. Looking at women's tennis, we find that the betting odds perform the best in terms of prediction accuracy and Brier score, while a simple weighted average of Elo and surface Elo outperforms the others in terms of model discrimination and expected return. The standard Elo ratings performed the best on calibration.

Figure 1: Forecasting performance (Wimbledon 2018 and 2019)

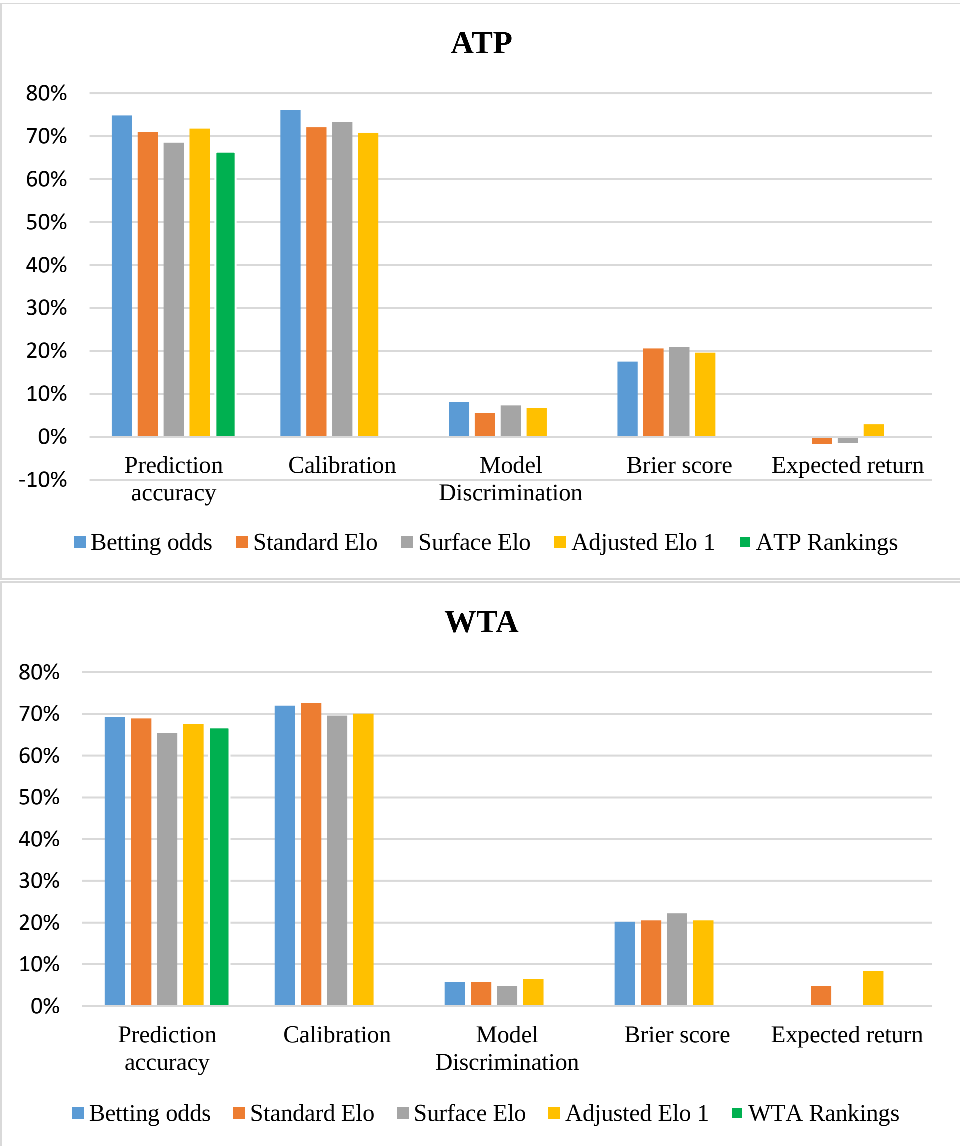


Table 4 summarizes the prediction by an adjusted Elo rating using Elo and surface Elo. Based on this search, almost all the forecasting measures are improved compared with the Elo rating itself. The optimal weights are different if we choose to maximize different forecasting measures. For example, if we use prediction accuracy as our target, we should set 87.0% on Elo rating for ATP but 75.6% on Elo rating for WTA.

It is noteworthy that the difference between the optimal weight on Elo for calibration is as pronounced as it can be. That the difference is so pronounced is perhaps a little surprising, but this is indeed what the data indicate. For the grass-courts that make up this Wimbledon data set, the surface is key in terms of calibration for men's tennis, while the opposite applies for women's tennis, where we can rely on standard Elo.

Table 4: Summary of prediction by weighted Elo and Grass surface ratings

Rating methods	Adjusted ATP Elo ratings 2	Adjusted WTA Elo ratings 2
Prediction accuracy	73.1%	71.4%
Optimal weight on Elo	87.0%	75.6%
Optimal weight on surface	13.0%	24.4%
Calibration	73.3%	72.7%
Optimal weight on Elo	0.0%	100.0%
Optimal weight on surface	100.0%	0.0%
Model discrimination	9.0%	6.9%
Optimal weight on Elo	40.5%	17.3%
Optimal weight on surface	59.5%	82.7%
Brier score	19.6%	20.3%
Optimal weight on Elo	56.7%	75.0%
Optimal weight on surface	43.3%	25.0%
Expected return	9.2%	13.3%
Optimal weight on Elo	85.9%	38.5%
Optimal weight on surface	14.1%	61.5%

As the role of betting odds is important in forecasting the performance, we construct another rating in the Equation (5) incorporating the betting odds. All the forecasting measures except Brier score have been improved with the betting odds. The corresponding optimal weights are shown in Table 5. For example, we should set the weight on Elo to be 1.9%, 0.0% on surface Elo and 98.1% on the betting odds to achieve the highest calibration in men's tennis.

Table 5[1]: Summary of prediction by weighted Elo, Grass surface ratings and betting odds

Rating methods	Adjusted ATP Elo ratings 3	Adjusted WTA Elo ratings 3
Prediction accuracy Optimal weight on Elo Optimal weight on surface Optimal weight on betting odds	74.8% Many combinations	71.4% Many combinations
Calibration	76.0%	72.7%
Optimal weight on Elo	1.9%	100.0%
Optimal weight on surface	0.0%	0.0%
Optimal weight on betting odds	98.1%	0.0%
Model discrimination	9.6%	7.8%

[1] It should be noted that there are no optimal weights related to the prediction accuracy reported for the Adjusted Elo ratings 3, as there are no unique solutions for the optimal weights. The only way to construct this adjusted Elo is through the weighted average of probabilities of winning (see Equation 5). We need to convert Elo, Elo surface and betting odds into probabilities first. Therefore, the adjusted Elo is a weighted average of winning probabilities. It is possible to calculate the maximum prediction accuracy but with many combinations of weights. This applies to the expected return as well.

Optimal weight on Elo Optimal weight on surface Optimal weight on betting odds	0.0% 52.4% 47.6%	24.8% 58.4% 16.8%
Brier score Optimal weight on Elo Optimal weight on surface Optimal weight on betting odds	18.3% 13.9% 0.0% 86.1%	20.1% 0.0% 20.1% 79.9%
Expected return Optimal weight on Elo Optimal weight on surface Optimal weight on betting odds	9.2% Many combinations	13.3% Many combinations

Tables 6 summarizes methods with the best forecasting performance. For men's tennis, betting odds are the best in terms of prediction accuracy, calibration, and Brier score. Adjusted Elo (a weighted composite of the betting odds, overall Elo and surface-specific Elo) is better in terms of model discrimination and expected return. For women's tennis, a weighted composite of the betting odds, overall Elo and surface-specific Elo performs best in terms of prediction accuracy, model discrimination, Brier score and expected return, while the standard Elo is best on calibration.

Table 6: Best performance of each method

Criteria	ATP		WTA	
	Best rating methods	**Weights**	**Best rating methods**	**Weights**
Prediction accuracy	Betting odds	NA	Adjusted Elo ratings 2 Adjusted Elo ratings 3	75.6% (Elo) 24.4% (surface) Many combinations
Calibration	Betting odds	NA	Standard Elo ratings	NA
Model discri-mination	Adjusted Elo ratings 3	0.0% (Elo) 52.4% (surface) 47.6% (betting odds)	Adjusted Elo ratings 3	24.8% (Elo) 58.4% (surface) 16.8% (betting odds)
Brier score	Betting odds	NA	Adjusted Elo ratings 3	0.0% (Elo) 20.1% (surface) 79.9% (betting odds)
Expected return	Adjusted Elo ratings 2 Adjusted Elo ratings 3	85.9% (Elo) 14.1% (surface) many combinations	Adjusted Elo ratings 2 Adjusted Elo ratings 3	38.5% (Elo) 61.5% (surface) many combinations

6.2 US Open 2019

Figure 2 shows the forecasting performance of US Open 2019. For men's tennis, we find that the betting odds outperform the other measures in terms of prediction accuracy and calibration. The standard Elo performs the best in terms of model discrimination, Brier score and expected return. Regarding women's tennis, a simple adjusted Elo rating performs better in terms of calibration and model discrimination, while standard Elo is better in terms of prediction accuracy and expected return. Betting odds has the lowest Brier score.

Figure 2: Forecasting performance (US Open 2019)

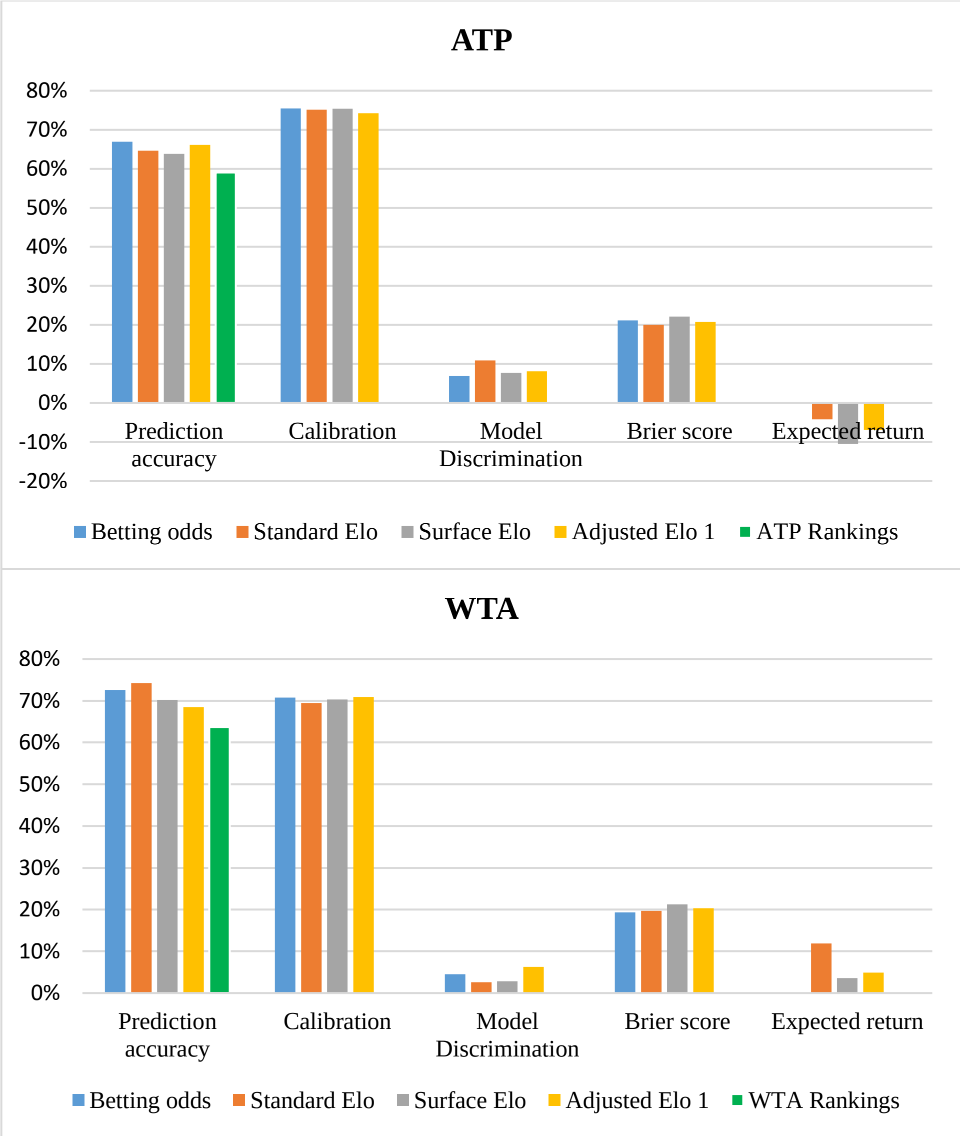


Table 7 summarizes the prediction by an adjusted Elo rating using Elo and surface Elo. We can see that almost all the forecasting measures are improved or at least the same as standard Elo rating in both ATP and WTA.

Table 7: Summary of prediction by weighted Elo and hard-court surface ratings

Rating methods	**Adjusted ATP Elo ratings 2**	**Adjusted WTA Elo ratings 2**
Prediction accuracy	66.9%	74.2%
Optimal weight on Elo	37.9%	89.2%
Optimal weight on surface	62.1%	10.8%
Calibration	75.7%	71.1%
Optimal weight on Elo	3.8%	41.2%
Optimal weight on surface	96.2%	58.8%
Model discrimination	10.9%	6.6%
Optimal weight on Elo	100.0%	41.2%
Optimal weight on surface	0.0%	58.8%
Brier score	20.0%	19.7%
Optimal weight on Elo	100.0%	100.0%
Optimal weight on surface	0.0%	0.0%
Expected return	3.2%	11.9%
Optimal weight on Elo	61.9%	100%
Optimal weight on surface	38.1%	0.0%

If we add betting odds, only model discrimination in ATP and Brier score in WTA are slightly improved (see Table 8).

Table 8: Summary of prediction by weighted Elo, hard-court surface ratings and betting odds

Rating methods	Adjusted ATP Elo ratings 3	Adjusted WTA Elo ratings 3
Prediction accuracy	66.9%	74.2%
Optimal weight on Elo	Many combinations	Many combinations
Optimal weight on surface		
Optimal weight on betting odds		
Calibration	75.6%	71.1%
Optimal weight on Elo	3.8%	41.2%
Optimal weight on surface	96.2%	58.8%
Optimal weight on betting odds	0.0%	0.0%
Model discrimination	10.94%	6.6%
Optimal weight on Elo	90.5%	41.2%
Optimal weight on surface	0.0%	58.8%
Optimal weight on betting odds	9.5%	0.0%
Brier score	20.0%	19.69%
Optimal weight on Elo	100.0%	0.0%
Optimal weight on surface	0.0%	18.5%
Optimal weight on betting odds	0.0%	81.5%
Expected return	3.2%	11.9%
Optimal weight on Elo	Many combinations	Many combinations
Optimal weight on surface		
Optimal weight on betting odds		

Tables 9 summarize methods with the best forecasting performance. The results are quite mixed. In general, adjusted Elo rating 2 and 3 are better than the other methods. The standard Elo is still the best in a couple of cases, such as Brier score in ATP and prediction accuracy in WTA.

Table 9: Best performance of each method

Criteria	**ATP**		**WTA**	
	Best rating methods	**Weights**	**Best rating methods**	**Weights**
Prediction accuracy	Betting odds	NA	Standard Elo ratings	NA
	Adjusted Elo ratings 2	37.9% (Elo) 62.1% (surface)	Adjusted Elo ratings 2	89.2% (Elo) 10.8% (surface)
	Adjusted Elo ratings 3	Many combinations	Adjusted Elo ratings 3	Many combinations
Calibration	Adjusted Elo ratings 2	3.8% (Elo) 96.2% (surface)	Adjusted Elo ratings 2	41.2% (Elo) 58.8% (surface)
Model discri-mination	Adjusted Elo ratings 3	90.5% (Elo) 0.0% (surface) 9.5% (betting odds)	Adjusted Elo ratings 2	41.2% (Elo) 58.8% (surface)

Brier score	Standard Elo ratings	NA	Adjusted Elo ratings 3	0.0 % (Elo) 18.5% (surface) 81.5% (betting odds)
Expected return	Adjusted Elo ratings 2 Adjusted Elo ratings 3	61.9% (Elo) 38.1% (surface) Many combinations	Standard Elo ratings Adjusted Elo ratings 3	NA Many combinations

6.3 Australian Open 2020

Figure 3 shows the forecasting performance for Australian Open 2020. For men's tennis, we find that the betting odds outperform the other metrics in terms of prediction accuracy, model discrimination and Brier score. In contrast, surface Elo outperforms the others in terms of calibration and expected return. For women's tennis, the betting odds exceed the other metrics in terms of prediction accuracy and Brier score. The standard Elo is the best in terms of calibration and model discrimination. The surface Elo outperforms the others in terms of expected return.

Figure 3: Forecasting performance (Australian Open 2020)

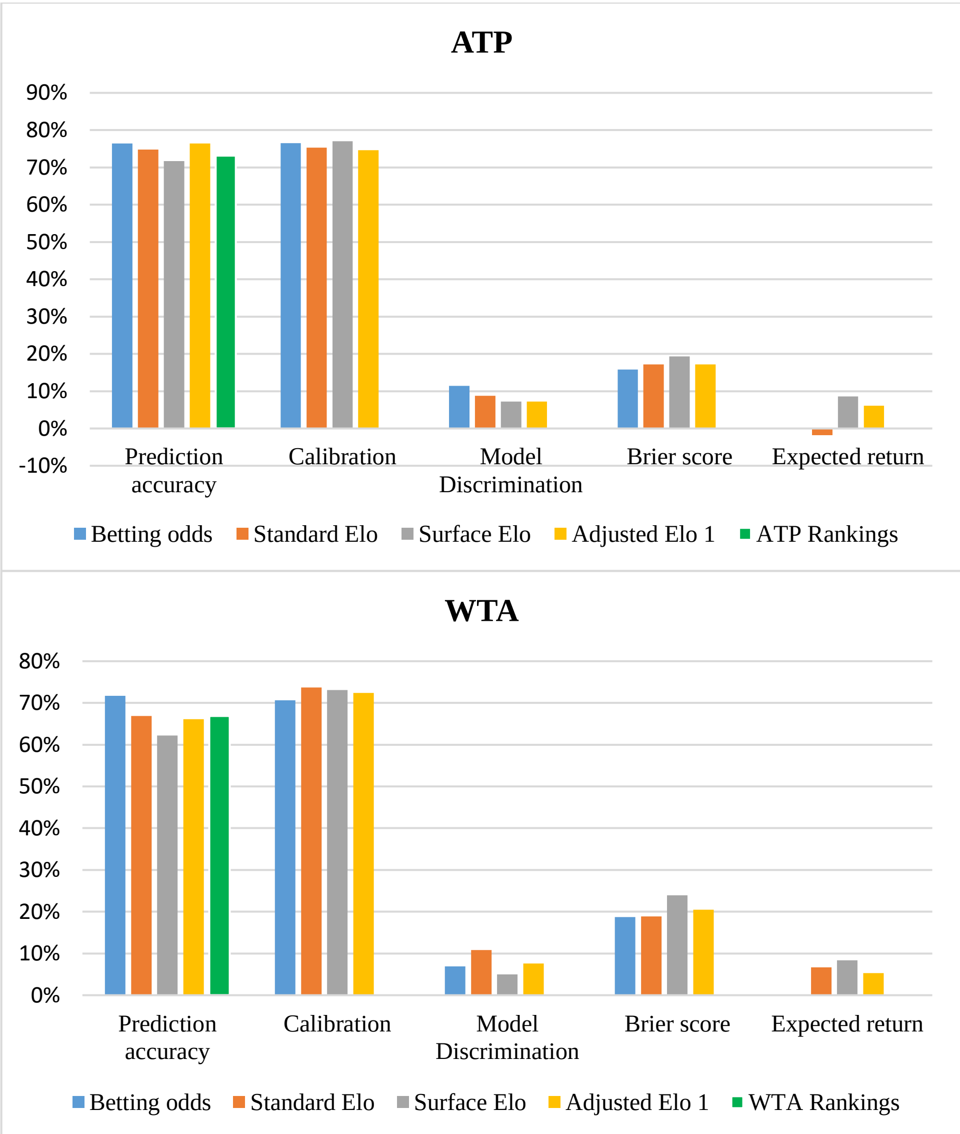


Table 10 summarizes the prediction by an adjusted Elo rating using Elo and surface Elo for the Australian Open. For both ATP and WTA, this adjusted Elo rating performs the best in terms of calibration and model discrimination. The standard Elo is the best in prediction accuracy and Brier Score in WTA. Surface Elo is still the best in the expected return of ATP.

Table 10: Summary of prediction by weighted Elo and hard-court surface ratings

Rating methods	Adjusted ATP Elo ratings 2	Adjusted WTA Elo ratings 2
Prediction accuracy	76.4%	66.9%
Optimal weight on Elo	50.0%	100.0%
Optimal weight on surface	50.0%	0.0%
Calibration	77.2%	74.0%
Optimal weight on Elo	30.8%	99.0%
Optimal weight on surface	69.2%	1.0%
Model discrimination	12.1%	11.4%
Optimal weight on Elo	31.7%	99.0%
Optimal weight on surface	68.3%	1.0%
Brier score	16.9%	18.9%
Optimal weight on Elo	76.8%	100.0%
Optimal weight on surface	23.2%	0.0%
Expected return	8.6%	12.9%
Optimal weight on Elo	0.0%	96.4%
Optimal weight on surface	100.0%	3.6%

The adjusted Elo rating 3 has been improved only in model discrimination in ATP (see Table 11). The other remaining methods in Figure 3 and Table 10 are still the best in ATP. In contrast, adjusted Elo 3 performed better in terms of calibration, model discrimination and expected return in WTA.

Table 11: Summary of prediction by weighted Elo, hard-court surface ratings and betting odds

Rating methods	Adjusted ATP Elo ratings 3	Adjusted WTA Elo ratings 3
Prediction accuracy	76.4%	71.7%
Optimal weight on Elo	Many combinations	Many combinations
Optimal weight on surface		
Optimal weight on betting odds		
Calibration	77.2%	74.1%
Optimal weight on Elo	30.8%	93.8%
Optimal weight on surface	69.2%	0.4%
Optimal weight on betting odds	0.0%	5.8%
Model discrimination	13.8%	12.2%
Optimal weight on Elo	44.4%	91.7%
Optimal weight on surface	3.9%	0.1%
Optimal weight on betting odds	51.7%	8.2%
Brier score	15.8%	18.7%
Optimal weight on Elo	0.0%	0.0%
Optimal weight on surface	0.0%	0.0%
Optimal weight on betting odds	100.0%	100.0%
Expected return	8.6%	13.6%
Optimal weight on Elo	Many combinations	Many combinations
Optimal weight on surface		
Optimal weight on betting odds		

Tables 12 summarize methods with the best forecasting performance. Betting odds perform best or joint best in forecasting prediction accuracy and Brier score in both men’s and women’s tennis. In contrast, adjusted Elo ratings 3 has the best or equivalently best performance in model discrimination and expected return. Surface Elo alone in ATP can generate the best expected return as well.

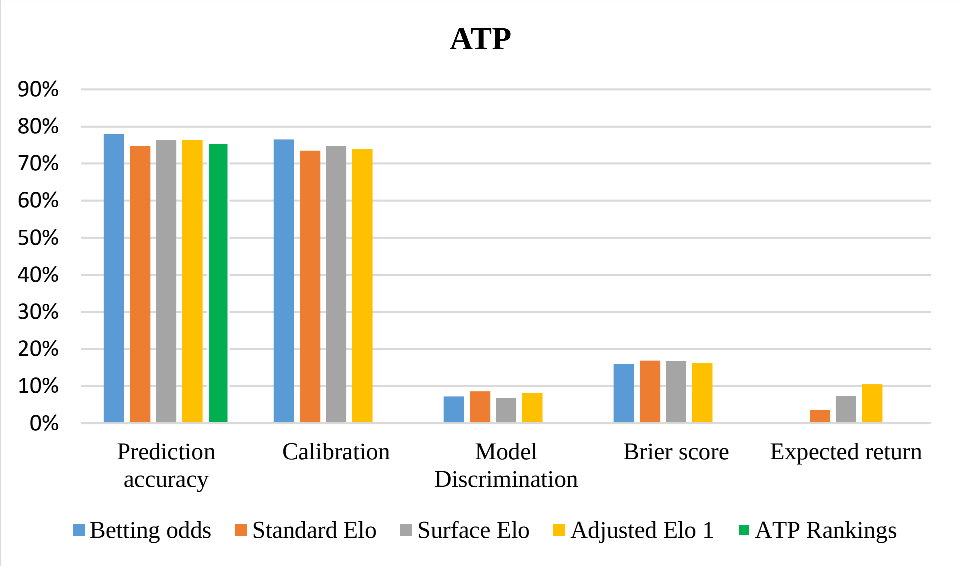
Table 12: Best performance in terms of each method

Criteria	**ATP**		**WTA**	
	Best rating methods	**Weights**	**Best rating methods**	**Weights**
Prediction accuracy	Betting odds Adjusted Elo ratings 1 Adjusted Elo ratings 3	NA 50.0% (Elo) 50.0% (surface) Many combinations	Betting odds Adjusted Elo ratings 3	NA Many combinations
Calibration	Adjusted Elo ratings 2	30.8% (Elo) 69.2% (surface)	Adjusted Elo ratings 3	93.8% (Elo) 0.4% (surface) 5.8% (betting odds)
Model discri-mination	Adjusted Elo ratings 3	44.4% (Elo) 3.9% (surface) 51.7% (betting odds)	Adjusted Elo ratings 3	91.7% (Elo) 0.1% (surface) 8.2% (betting odds)
Brier score	Betting odds	NA	Betting odds	NA
Expected return	Surface Elo Adjusted Elo ratings 3	NA Many combinations	Adjusted Elo ratings 3	Many combinations

6.4 French Open 2019

Figure 4 shows the forecasting performance for French Open 2019. In general, betting odds perform better in prediction accuracy and Brier score. The standard Elo is the best in terms of model discrimination in ATP, and prediction accuracy and calibration in WTA. A simply adjusted Elo is the best in respect of expected return in ATP and prediction accuracy in WTA.

Figure 4: Forecasting performance (French Open 2019)



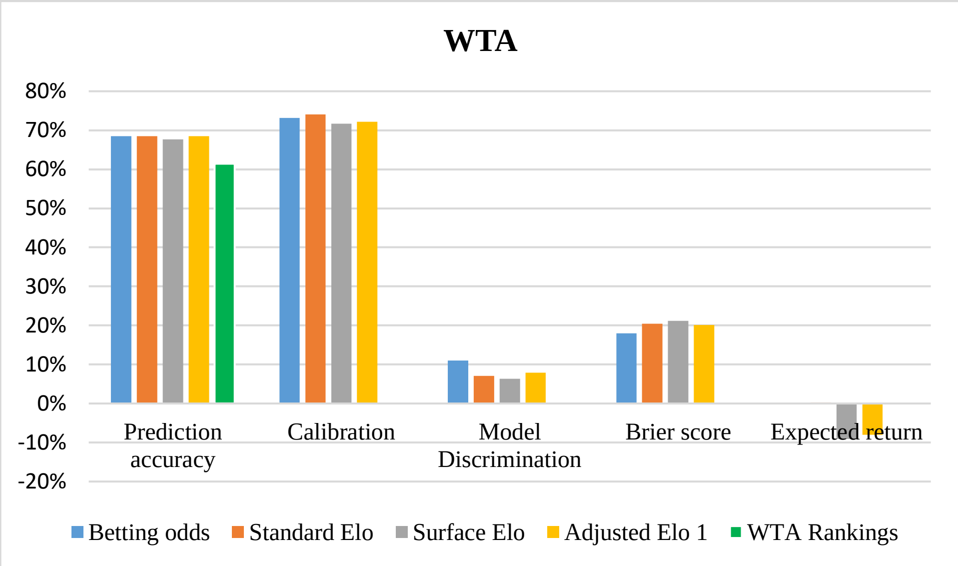


Table 13 summarizes the prediction by an adjusted Elo rating using Elo and surface Elo for 2019 French Open. For both ATP and WTA, this adjusted Elo rating performs the best in terms of model discrimination and expected return in ATP, and prediction accuracy and calibration in WTA.

Table 13: Summary of prediction by weighted Elo and clay surface ratings

Rating methods	Adjusted ATP Elo ratings 2	Adjusted WTA Elo ratings 2
Prediction accuracy	78.0%	70.1%
Optimal weight on Elo	45.1%	38.9%
Optimal weight on surface	54.9%	61.1%
Calibration	74.9%	74.2%
Optimal weight on Elo	2.4%	95.4%
Optimal weight on surface	97.6%	4.6%
Model discrimination	8.9%	9.3%
Optimal weight on Elo	67.6%	64.1%
Optimal weight on surface	32.4%	35.9%
Brier score	16.3%	20.08%
Optimal weight on Elo	51.5%	64.3%
Optimal weight on surface	48.5%	35.7%
Expected return	12.2%	0.3%
Optimal weight on Elo	33.6%	100.0%

Optimal weight on surface	66.4%	0.0%

All the forecasting measures except prediction accuracy and calibration have been improved with the betting odds (see Table 14) where we can see that the betting odds play an essential role in forecasting these measures.

Table 14: Summary of prediction by weighted Elo, clay surface ratings and betting odds

Rating methods	**Adjusted ATP Elo ratings 3**	**Adjusted WTA Elo ratings 3**
Prediction accuracy Optimal weight on Elo Optimal weight on surface Optimal weight on betting odds	78.0% Many combinations	66.9% Many combinations
Calibration Optimal weight on Elo Optimal weight on surface Optimal weight on betting odds	76.5% 0.0% 0.0% 100.0%	74.2% 94.4% 0.0% 5.6%
Model discrimination Optimal weight on Elo Optimal weight on surface Optimal weight on betting odds	10.0% 47.5% 0.0% 52.5%	11.6% 49.7% 0.0% 50.3%
Brier score Optimal weight on Elo Optimal weight on surface Optimal weight on betting odds	15.5% 0.0% 15.1% 84.9%	18.0% 0.0% 0.0% 100.0%
Expected return Optimal weight on Elo Optimal weight on surface Optimal weight on betting odds	13.7% Many combinations	0.3% Many combinations

Tables 15 summarize methods with the best forecasting performance for French Open 2019. This adjusted Elo with the betting odds outperforms the other metrics in terms of model discrimination and expected return in both ATP and WTA. It also performs the best or jointly best in respect of prediction accuracy and Brier score in men's tennis. Betting odds are the best or joint best in terms of prediction accuracy and calibration in ATP and Brier score in WTA.

Table 15: Best performance in terms of each method

Criteria	**ATP**		**WTA**	
	Best rating methods	**Weights**	**Best rating methods**	**Weights**
Prediction accuracy	Betting odds Adjusted Elo ratings 2 Adjusted Elo ratings 3	NA 445.1% (Elo) 54.9% (surface) Many combinations	Adjusted Elo ratings 2	38.9% (Elo) 61.1% (surface)
Calibration	Betting odds	NA	Adjusted Elo ratings 2	95.4% (Elo) 4.6% (surface)
Model discri-mination	Adjusted Elo ratings 3	47.5% (Elo) 0.0% (surface)	Adjusted Elo ratings 3	49.7% (Elo) 0.0% (surface)

		52.5% (betting odds)		50.3% (betting odds)
Brier score	Adjusted Elo ratings 3	0.0% (Elo) 15.1% (surface) 84.9% (betting odds)	Betting odds	NA
Expected return	Adjusted Elo ratings 3	Many combinations	Adjusted Elo ratings 3	Many combinations

6.5 Differences of forecasting performance between higher-ranked and lower-ranked players

It is interesting to see if there is any difference in predicting the matches between higher Elo-ranked players and lower Elo-ranked players. As most of the forecasting performances are calculated by matches rather than players, we split the data into matches which include higher-ranked players and matches consisting of lower-ranked players in terms of standard Elo ranking. We use data from all matches played in our Wimbledon data sets. Higher-ranked players are defined as those in the top 30 Elo. This dividing line serves to split the sample relatively evenly.

Figure 5 and Figure 6 show the forecasting performance for the higher-ranked and lower-ranked group, respectively. On prediction accuracy, calibration, model discrimination and Brier score, the higher-ranked category performs better. In terms of expected return, the lower-ranked category performs better in almost all cases.

Figure 5: Forecasting performance (high-ranked group)

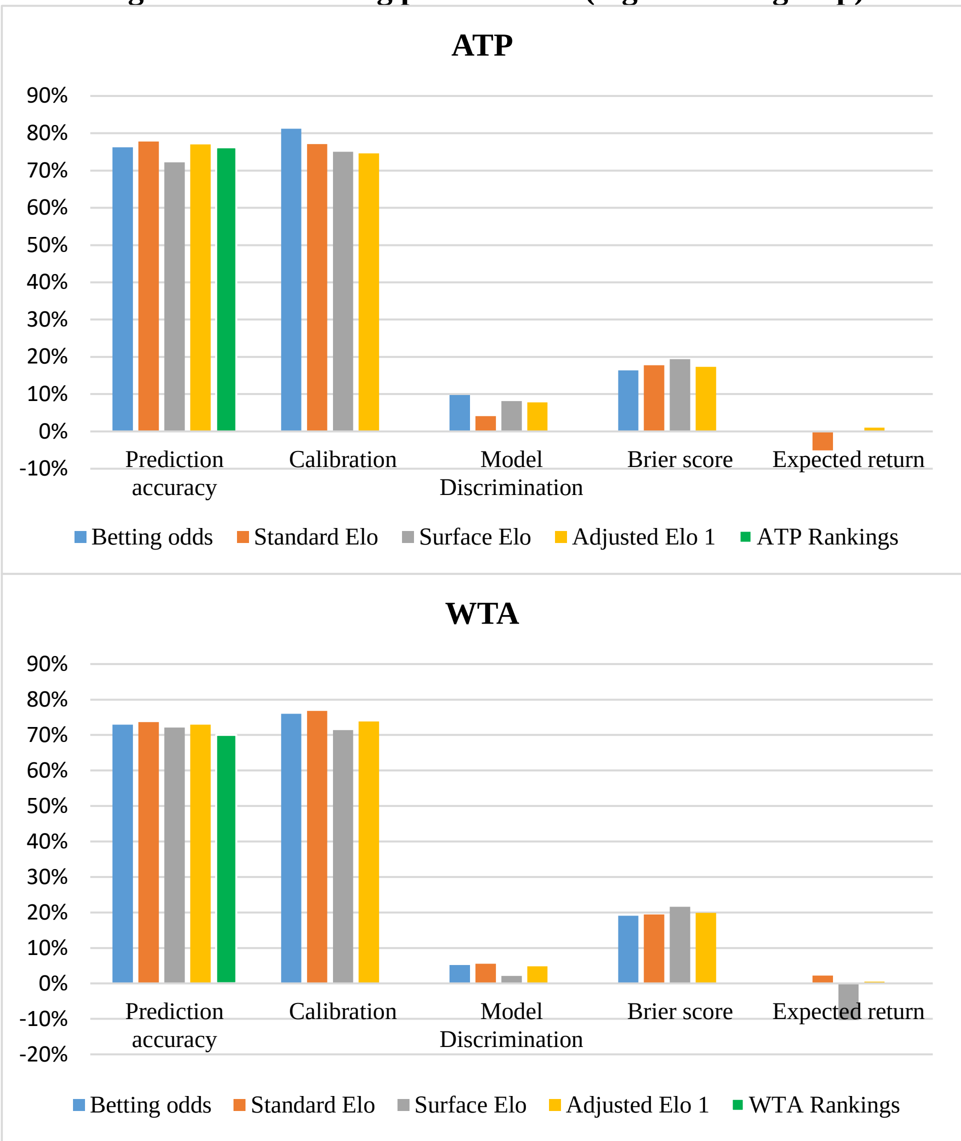
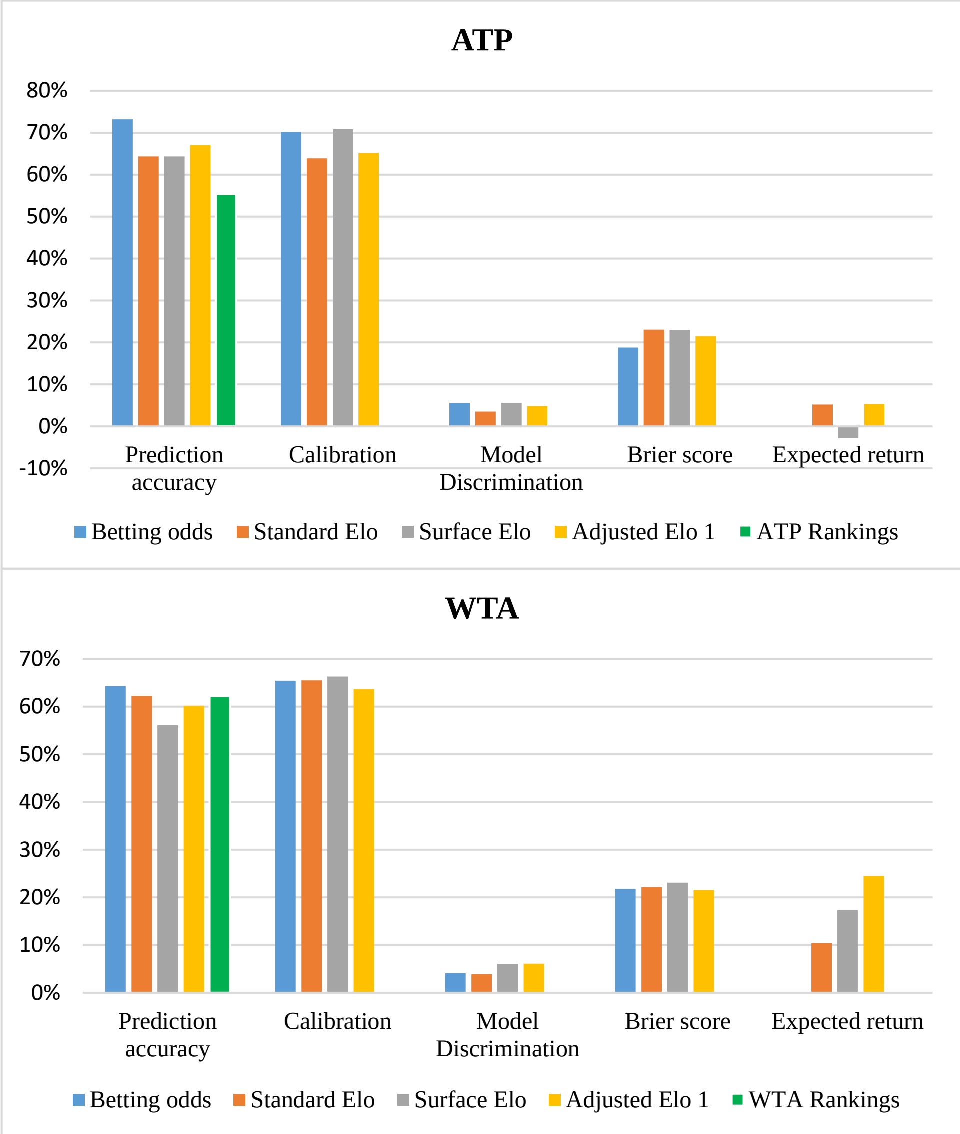


Figure 6: Forecasting performance (low-ranked group)



7. Summary of results

The measures we use are the betting odds, the official tennis rankings and the overall Elo ratings, as well as explicit use of both surface-specific Elo ratings and of weighted composites of Elo and surface Elo ratings, including and excluding the betting odds. The performance indicators used are prediction accuracy, calibration, model discrimination, Brier score and expected return. We perform the analysis for both men's tennis and women's tennis.

For men's tennis at grass-court Wimbledon, we find that betting odds perform best in terms of prediction accuracy, calibration, and Brier score. Adjusted Elo (a weighted composite of the betting odds, overall Elo and surface-specific Elo) is better in terms of model discrimination and expected return. For women's tennis, a weighted composite of the betting odds, overall Elo and surface-specific Elo performs best in terms of prediction accuracy, model discrimination, Brier score and expected return, while the standard Elo is best for calibration.

For men's tennis at the hard-court US Open, we find that the betting odds outperform the other measures in terms of prediction accuracy and calibration. The standard Elo performs the best in terms of model discrimination, Brier score and expected return. Regarding women's tennis, a simply adjusted Elo rating performs better in terms of calibration and model discrimination, while standard Elo is better in terms of prediction accuracy and expected return. Betting odds has the lowest Brier score.

For men's tennis at the hard-court Australian Open, we find that the betting odds outperform the other measures in terms of prediction accuracy, model discrimination and Brier score. In contrast, surface Elo outperforms the others in terms of calibration and expected return. For women's tennis, the betting odds exceed the other metrics in terms of prediction accuracy and Brier score. The standard Elo performs best on calibration and model discrimination. The surface Elo outperforms the others in terms of expected return.

At the clay-court French Open, the adjusted Elo incorporating the betting odds outperforms the other measures in terms of model discrimination and expected return for both men's tennis and women's tennis. The betting odds perform best or joint best in terms of prediction accuracy and calibration in men's tennis and the Brier score in women's tennis.

In our selected data sets, we find that matches including the category of higher-ranked (top 30 Elo) players performed best on all measures except expected return.

8. Conclusion

This paper seeks to compare and evaluate the performance of five different measures for forecasting men's and women's professional tennis matches. We use data derived from every match played at the 2018 and 2019 Wimbledon tennis championships, the 2019 French Open, the 2019 US Open and the 2020 Australian Open. We use the betting odds, the official tennis rankings, the overall Elo ratings, the surface-specific Elo ratings and a composite of some of the above. The Elo rating system is a method of ranking players based on their past matches, weighted by the ratings of the players they competed against. The performance indicators we use are prediction accuracy, calibration, model discrimination, Brier score and expected return.

We find that the betting odds perform well on a number of performance indicators across all tournaments, while standard Elo (especially for women's tennis) and surface-adjusted Elo (especially for men's tennis) also perform well on other performance indicators. For all but the hard-court surfaces, a forecasting model which incorporates the betting odds tends to perform particularly well on some performance indicators.

Consistently, however, we find that the official ranking system (where it could be compared with other forecasting metrics, including notably Elo-based ratings) proved to be a relatively poor measure of likely current performance (see also Reid et al., 2010).

We also find that our adjusted Elo rating is a better predictor for higher-ranked players (top 30) in terms of every measure except for expected return.

We can conclude that the betting odds, or an adjusted Elo measure which incorporates the betting odds, performs best or joint best on most forecasting measures at Wimbledon and the French Open, which are grass-court and clay-court respectively. For men's tennis and women's tennis at the US Open and Australian Open, the betting odds perform well on most performance indicators, while standard Elo and a simple surface-adjusted Elo performs best on others.

Importantly, the way in which the rankings are constructed is also a vital consideration in the pay structure of competitors, as these rankings determine tournament entry qualification, seedings, and associated prize money and sponsorship. The uses of Elo-based methodologies can and have also been used outside of the competitive arena to measure performance. In particular, the Elo ratings methodology can be used in education by interpreting a solution attempt as a match between a student and an item (e.g. Mangaroska et al., 2019). Other uses of Elo-based systems include the use of an Elo rating algorithm to calculate medical website 'credibility' values, in soft biometrics, computer vision, and a variety of matchmaking applications.

More specifically for the case of tennis, the conclusions of this paper complement and build upon those of earlier studies, notably Kovalchik (2016) – see also Kovalchik and Reid (2019) - who studied the predictive ability of previously published tennis prediction models.

In summary, the findings of this paper add further weight to the case for a wider use of Elo-based approaches within sports forecasting (including weighted composite measures) as well as arguably within the player rankings methodologies.

References

Angelini, G. and De Angelis, L., 2019. Efficiency of online football betting markets. *International Journal of Forecasting*, 35 (2), 712-721.

Asch, P., Malkiel, B.G. and Quandt, R.E. 1984. Market efficiency in racetrack betting. *Journal of Business*, 57. 165-175.

ATP (Association of Tennis Professionals) ATP Rankings. Available at: www.apttour.com/en/rankings/singles/. Last accessed on 2 February 2020. Bolton, R.N. and Chapman, R.G., 1986. Searching for positive returns at the track: A multinomial logit model for handicapping horse races. *Management Science*, 32 (8), 1040-1060.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78 (1), 1-3.

Carbone, J., Corke, T. and Moisiadis, F., 2016. The Rugby League Prediction Model: Using an Elo-based approach to predict the outcome of National Rugby League (Nrl) matches. *International Educational Scientific Research Journal*, 27 (9), 26-30.

Clarke, S., Kovalchik, S. and Ingram, M., 2017. Adjusting bookmaker's odds to allow for overround. *American Journal of Sports Science*, 5 (6), p. 45.

Dingle, N., Knottenbelt, W. and Spanias, D., 2012. On the (page) ranking of professional tennis players. *European Workshop on Performance Engineering*, 237-247. Springer, Berlin, Heidelberg.

Easton, S. and Uylangco, 2010. Forecasting outcomes in tennis matches using within-match betting markets. *International Journal of Forecasting*, 26 (3), 564-575.

Elo, A.E., 1978. *The rating of chessplayers, past and present*. Arco Pub.

Flashscore (2018). Available at: https://www.flashscore.com/tennis/. Last accessed on 15 July, 2018.

Flashscore (2019). Available at: https://www.flashscore.com/tennis/. Last accessed on 15 October, 2019.

Flashscore (2020). Available at: https://www.flashscore.com/tennis/. Last accessed on 28 May, 2020.

Graham, I. and Stott, H. 2008. Predicting bookmaker odds and efficiency for UK football. *Applied Economics*, 40 (1), 99-109.

Hvattum, L.M. and Arntzen, H., 2010. Using Elo ratings for match result prediction in association football. *International Journal of Forecasting*, 26 (3), 460-470.

Irons, D.J., Buckley, S. and Paulden.T., 2014. Developing an improved tennis ranking system. *Journal of Quantitative Analysis in Sports*, 10 (2), 109-118.

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

Kovalchik, S.A. and Reid, M. 2019. A Calibration Method with Dynamic Updates for Within-Match Forecasting of Wins in Tennis. *International Journal of Forecasting*, 35 (2), 756-766.

Leitner, C., Zeileis, A. and Hornik, K., 2009. Is Federer stronger in a tournament without Nadal? An evaluation of odds and seedings for Wimbledon 2009. *Austrian Journal of Statistics*, 38 (4), 277-286.

Leitner, C., Zeileis, A. and Hornik, K., 2010. Forecasting sports tournaments by ratings of (prob)abilities: A comparison for EURO 2008. *International Journal of Forecasting*, 26 (3), 471-481.

Lisi, F., 2017. Tennis betting: can statistics beat bookmakers? *Electronic Journal of Applied Statistical Analysis*, 10 (3), 790-808.

Mangaroska, K., Vesin, B. and Giannakos, M. 2019. Elo-Rating Method: Towards Adaptive Assessment in E-Learning. *IEEE*, 1-3.

Morris, B., Bialik, C. and Boice, J., 2016. How We're Forecasting the 2016 U.S. Open. FiveThirtEight.com. August 28. Available at: https://fivethirtyeight.com/features/how-were-forecasting-the-2016-us-open/. Last accessed on 15 October, 2019.

Oddschecker, 2018. Available at: www.oddschecker.com Last accessed on 15 July, 2018.

Oddschecker, 2019. Available at: www.oddschecker.com Last accessed on 8 September, 2019.

Oddschecker, 2020. Available at: www.oddschecker.com Last accessed on 2 February, 2020.

Pencina, M.J., D'Agostino, R.B. and Vasan, R.S., 2008. Evaluating the added predictive ability of a new marker: From area under the Roc curve to reclassification and beyond. *Statistics in medicine*, 27 (2), 157-172.

Pope, P.F. and Peel, D.A., 1989. Information, prices and efficiency in a fixed-odds betting market. *Economica*, 323-34.

Reade, J.J., Singleton, C. and Vaughan Williams, L., 2020. Betting markets for English Premier League results and scorelines: evaluating a forecasting model. *Economic Issues,* 25 (1), 85-106.

Reid, M., McMurtrie, D. and Crespo, M., 2010. The relationship between match statistics and top 100 ranking in professional men's tennis. *International Journal of Performance Analysis in Sport*, 10 (2), 131-138.

Ryall, R. and Bedford, A., 2010. An optimized ratings-based model for forecasting Australian Rules football. *International Journal of Forecasting*, 26 (3), 511-517.

Silver, N. and Fischer-Baum, R., 2015. How We Calculate NBA Elo Ratings. FiveThirtyEight.com. May 21. Available at: https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/

Snyder, W. W., 1978. Horse racing: testing the efficient markets model. *Journal of Finance*, 33, 1109-1118.

Sonas, F., 2002. The Sonas rating formula: Better than Elo? Chessbase news. Retrieved from: http://chessbase.com/newsdetail.asp?newsid=562 Last accessed on 30 May, 2020.Spann, M. and Skiera, B., 2009. Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, 28 (1), 55-72.

Stekler, H., Sendor, D. and Verlander, R. 2010. Issues in sports forecasting. *International Journal of Forecasting*, 26 (3), 606-621.

Tennis Abstract, 2018-2020a. Current Elo ratings for the ATP tour. Available at: www.tennisabstract.com/reports/atp_elo_ratings.html. Last accessed on 2 February, 2020.

Tennis Abstract, 2018-2020b. Current Elo ratings for the WTA tour. Available at: www.tennisabstract.com/reports/wta_elo_ratings.html. Last accessed on 2 February, 2020.Tetlock, P.E. and Gardner, D. 2015. Superforecasting: The Art and Science of Prediction. Crown Publishers.

Vaughan Williams, L. and Stekler, H. 2010. Sports forecasting. *International Journal of Forecasting*, 26 (3), 445-447.

Vaughan Williams, L. 2005, ed. Information Efficiency in Financial and Betting Markets. Cambridge: Cambridge University Press.

Vergin, R.C. and Scriabin, M., 1978. Winning strategies for wagering on national football league games. *Management Science*, 24 (8), pp. 809-818.

Walkofmind, 2019. Elo rating vs. winning probabilities. Available at: https://www.walkofmind.com/prgramming/chess/elo.htm. Last accessed on 15 October, 2019.

WTA (Women's Tennis Association) Rankings. Available at: www.wtatennis.com/rankings. Last accessed on 2 February 2020.