Proceedings of Machine Learning Research vol 245:1–11, 2024 2024 International Conference on Machine Learning and Intelligent Computing

Forecasting Tennis Player Matches Based on Machine Learning

Rui Bai
Xiamen University Malaysia, Jalan Sunsuria, Bandar Sunsuria, 43900 Sepang, Selangor, Malaysia
Kar Hing Chong
Xiamen University Malaysia, Jalan Sunsuria, Bandar Sunsuria, 43900 Sepang, Selangor, Malaysia
Haoyuan Li
Xiamen University Malaysia, Jalan Sunsuria, Bandar Sunsuria, 43900 Sepang, Selangor, Malaysia
Jia Yew Teh* JIAYEW.TEH@XMU.EDU.MY
Xiamen University Malaysia, Jalan Sunsuria, Bandar Sunsuria, 43900 Sepang, Selangor, Malaysia

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

This paper aims to highlight the extensive potential of analytics with the use of machine learning to improve sports modelling. We propose a supervised machine learning approach to further extend the optimization of machine learning in predicting the flow of points in tennis matches. Using data sourced from the 2023 Wimbledon Gentlemen's singles matches, we used Grey Relational Analysis and membership functions from fuzzy set theory to extract and rank 7 features that exhibit impactful ties with the player's match outcome, which includes player's serve status, games won in current set, ranking difference, distance covered, serve speed, previous victory status, and unforced error, respectively. We implemented these features to build 3 supervised models and compare their predictive performances, namely K-Nearest Neighbours, XGBoost and Logistic Regression. We adopted a train test split measure of 300 training sets and 100 testing sets. Using performance metrics such as confusion matrices, ROC curves, F1, Precision, Recall, and Accuracy scores, we constructed a scoring table to rank implemented models. Our results demonstrated that XGBoost exhibited the most significant predictive performance, followed by KNN and Logistic Regression. 5-Fold cross-validation feature stability and sensitivity analysis suggests that the feature space created is robust and stable where features are not easily subject to change in short-term predictions.

Keywords: Supervised Learning, Point Forecast, Fuzzy Set, Grey Relational Analysis, K-Nearest Neighbours, XGBoost, Logistic Regression

1. Introduction

As tennis becomes increasingly popular amongst other sports, players and coaches show great interest in analyzing tennis matches in hopes of finding patterns to gain an advantage when facing their next opponent. With diverse data gathered from past major tennis tournaments, machine learning becomes a suitable candidate for efficient analysis of historical data to provide relatively accurate predictions. Hence, we propose a structured framework using feature engineering and supervised model training to analyze historic matches, mainly the 2023 Wimbledon Gentlemen's singles matches. Substantial studies have presented diverse methods to model professional tennis matches as the sport gains popularity. Previous research has utilized machine learning techniques by optimizing logistic regression and artificial neural networks to predict match outcomes (Sipko and Knottenbelt, 2015). Similarly, studies from Fernandes (2017) as well as Candila and Palazzo (2020) introduced ANN for tennis modelling. Wilkens (2021) implemented various models to compare and evaluate prediction outcomes.

We attempt to implement an alternate tool to perform feature engineering in tennis analytics by extracting important features through Gray Relational Analysis (GRA) in contrast to conventional statistical techniques (Song and Shepperd, 2011). Moreover, we extend existing frameworks that map psychological pressure from ranking differences between players by constructing membership functions using concepts from fuzzy set theory. After extracting features from the provided data, we fit and compare multiple models to select the model best fit for forecasting tennis matches. Finally, we wish to expand this research by including modern algorithms in our analysis, such as XGBoost, which is a variant of Gradient-Boosted Decision Tree (GBDT) algorithms.

2. Feature Engineering using Grey Relational Analysis

2.1. Related Features

Feature engineering transforms raw data to optimize training speed and enhance prediction accuracy. Statistical interpretation becomes more meaningful and straight-forward as it provides variables that are self-explanatory and less abstract.

Through observations, we noticed that the scores of each game fluctuated in respect to time. Once this connection is verified, we begin to implement Grey Relational Analysis (GRA) to filter out features that will be used as additional inputs for super-vised algorithms (Sallehuddin et al., 2008). Specifically, Song and Shepperd (2011) suggested that GRA provides substantial variability in extracted features which indicates major correlations to the subject of analysis. It handles well with complex data with high dimensions and can generate competent results with insufficient information. This approach allows us to precisely describe how these relevant features correlate with hidden patterns that describe the flow of points. Our process can be summarized in Figure 1.

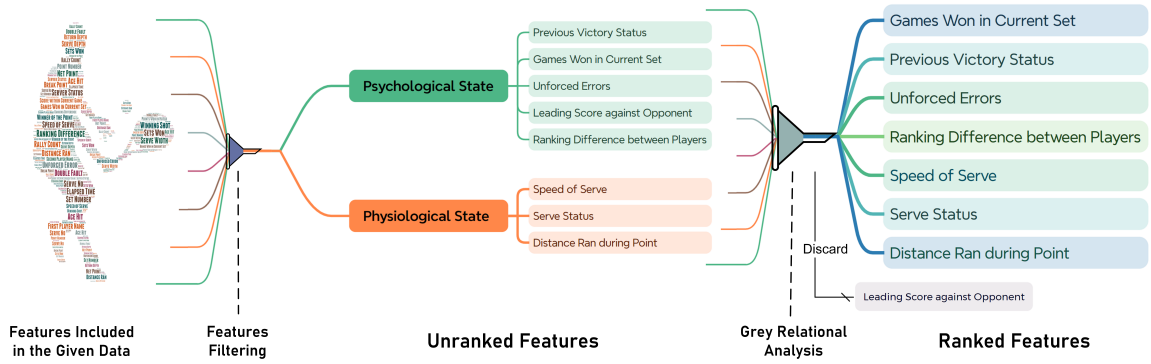

Figure 1: Big picture of filtering and ranking parameters.

Table 1 visualizes the series of features that are closely correlated to whether a point has been won by a player (labelled as Y). These features are considered with-in the official rules of tennis matches.

2.2. Establishing Membership Function between Player Rankings

We postulate that ranking differences are open to ambiguous results of how psycho-logical pressure affects players in games, denoted by x_6 as one of the mentioned features. Chiang and Denes (2023) provided a simple non-linear approach to map ranking differences to one of the features in their

Table 1: Notations of Filtered but Unranked Features.

Symbols	Definitions
x_1	Player's serve status
x_2	Player's previous victory status
x_3	Games won by player in current set
x_4	Speed of serve (miles per hour, mph)
x_5	Player's distance covered during point (meters)
x_6	Ranking difference between two players
x_7	Unforced error of player
x_8	Leading score of player against another player
Y	Player's current victory status

model. We extend this framework by implementing fuzzy sets and assign degrees of memberships to each ranking difference. This is extremely useful for circumstances when the level of psychological pressure is not discrete, instead it is subjective to ambiguous boundaries. We can evaluate pressure statistically through fuzzy [1] relationships that classical sets otherwise struggle to interpret (Zimmermann, 2011). Membership Functions [2] were assigned to further analyze how big of a difference in ranking relates to psychological pressure between players. Ranking data was gathered through the ATP Tour Website (ATP, 2024).

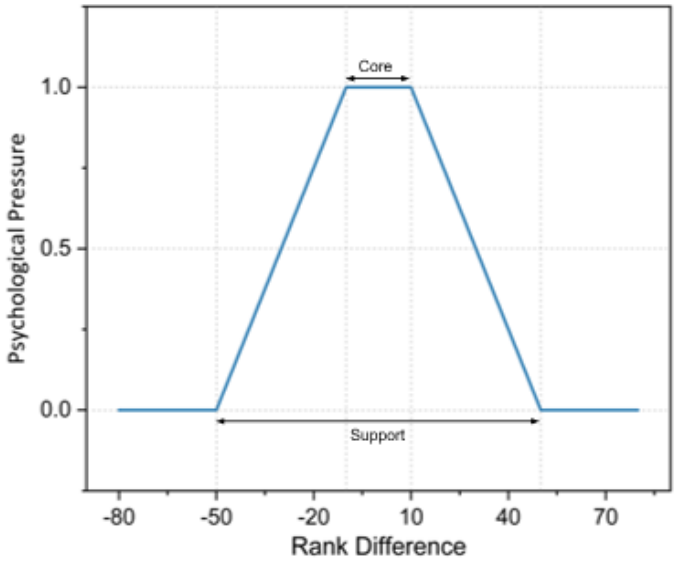

Figure 2: Membership function of psychological pressure with ranking differences.

1. develops a quantitative connection between variables when information on categorizing them is unknown or uncertain
2. Member of a fuzzy set with a degree of membership between [0, 1]

We follow the definition of fuzzy set theory to define how we relate differences in ranking to psychological pressure using this expression.

$$\mathrm{A} = \{(x, \mu_{\mathrm{A}}(x)) \mid x \in U\} \Rightarrow \mu\mathrm{A}(x) : \mathrm{U} \mapsto [0, 1] \tag{1}$$

where x denotes the difference in ranking, A the set that describes psychological pressure, $\mu\mathrm{A}(x)$ the membership function of element x in set A which maps the Universe of Discourse U to interval [0, 1]. Figure 2 illustrates a trapezoidal function that describes how closely relevant is the fuzzy relationship. $Support(A)$ explains that the psychological pressure is bounded by ranking differences in interval [-50, 50], this implies that any difference under 50 shows some form of psychological impact on stress. On the other hand, $Core(A)$ shows that differences under 10 rankings have a direct impact on psychological pressure.

$$u_A(x) = \begin{cases} 0 & \text{if } x < -50 \\ \frac{x+50}{40} & \text{if } -50 \leq x \leq -10 \\ 1 & \text{if } -10 \leq x \leq 10 \\ \frac{50-x}{40} & \text{if } 10 \leq x \leq 50 \\ 0 & \text{if } x > 50 \end{cases} \tag{2}$$

2.3. Ranking Data with Grey Relational Grade

Data processing was done prior to this step. Simply assigning quantitative identification numbers to label data in nominal forms is considered improper practice of data handling. Hence, we rigorously perform One-Hot Encoding [3] to avoid certain side effects such as overfitting and sparse data. Transformation is visualized in Table 2. Obtained features were standardized for further calculations.

Table 2: Transformation of Categorical Data using One-Hot Encoding.

Serve Width		Body Center	Body	Body Wide	Wide	Center
Body Center		1	0	0	0	0
Body		0	1	0	0	0
Body Wide	$\rightarrow$	0	0	1	0	0
Wide		0	0	0	1	0
Center		0	0	0	0	1

Grey Relational Grade (GRG) is obtained by taking the average of Grey Relational Coefficients (GRC) distributed across comparability sequences in relation to time, which is the following expression.

$$\gamma_i = \frac{1}{n} \sum_{k=1}^{n} \zeta_i(k) \tag{3}$$

Implementing this method allows us to rank as well as filter unnecessary features, as illustrated in Table 3.

3. Labeling technique for conversion of categorical values to numerical values in Machine Learning models

Table 3: Ranked Features.

Symbols	GRG	Ranking
x_1	0.940	1
x_3	0.916	2
x_6	0.914	3
x_5	0.913	4
x_4	0.912	5
x_2	0.890	6
x_7	0.866	7
x_8	0.707	8

From the results we obtained, we can clearly see that the serving status of the player is the leading factor that impacts the flow of points. x_8 is discarded since the importance of impact is relatively not significant.

2.4. Model Building

Evidently, tennis matches are sets of historical data containing various data that is already gathered beforehand. This allows the suitable use of supervised algorithms where we clearly know what the output value is when we train our data, which is the outcome of the match. From the Wimbledon 2023 Gentlemen Singles tournament dataset (COMAP, 2024), we adopted the train-test split approach where 300 matches were used as training data, while 100 matches were used as testing sets. Furthermore, we selected three widely used supervised algorithms to predict how points progress throughout the match.

K-Nearest Neighbours: KNN is a supervised, non-parametric machine learning model that groups data with similar characteristics, providing predictions based on the nearest Euclidean distance [4] (Zhang et al., 2018). It classifies data points based on the majority of k neighbouring types with k as a small positive integer, usually odd.

XGBoost: Extreme Gradient Boosting (XGBoost) is a variant of Gradient-Boosted Decision Tree (GBDT) algorithms which builds off from features of training data, creates and evaluates decision trees to estimate results (Santhanam et al., 2017). XGBoost handles classifiers by correcting mistakes sequentially from previous trees. Another advantage suggests that the algorithm is scalable towards extremely large datasets (Chen and Guestrin, 2016).

Logistic Regression: This type of regression algorithm maps inputs consisting of real values into probabilities between 0 and 1 using sigmoid functions. The estimation of outcomes is calculated through maximum likelihood to find the best parameter (David W. Hosmer Jr., 2013). Such parameters give better statistical interpretability on the importance of each feature obtained through GRA.

For match predictions, we chose to specifically use binary classification to label wins and losses with discrete values of 0 and 1 based on the player's current win status Y. In the case of logistic regression, we map probabilities less than 0.5 to be 0 while probabilities larger than 0.5 to be 1.

4. Note that not every classification task uses Euclidean distance, alternative metrics are available

2.5. Model Results and Interpretation

Figure 3 visualizes confusion matrices based on KNN, XGBoost, and logistic regression. Figure 4 presents the ROC curves evaluated for classification performance.

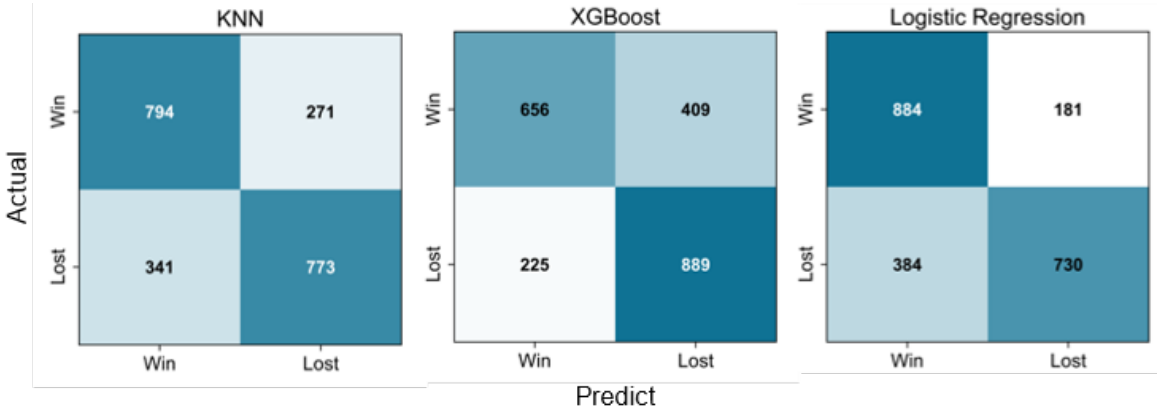

Figure 3: Confusion Matrices for KNN, XGBoost, Logistic Regression.

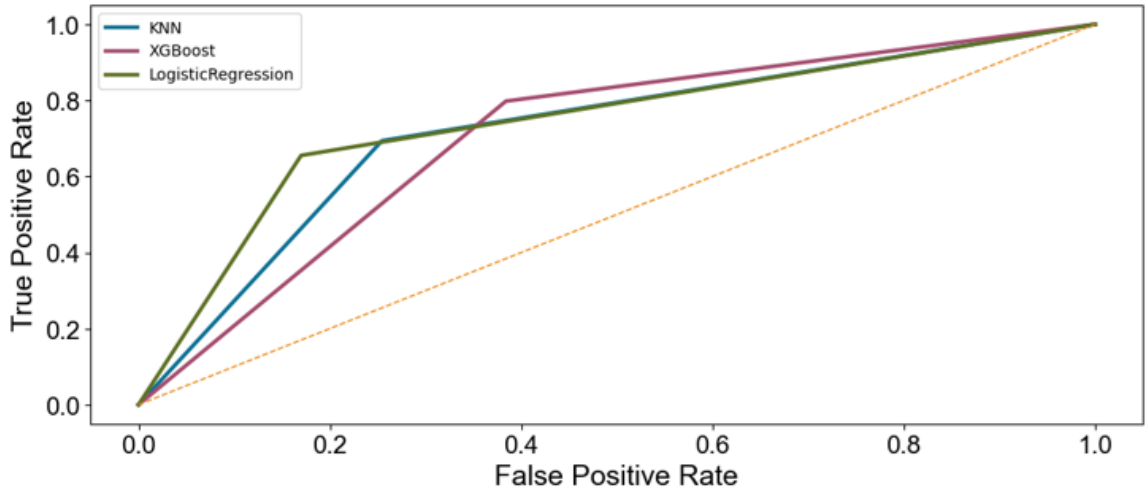

Figure 4: ROC Curve for KNN, XGBoost, Logistic Regression.

Table 4 illustrates a brief summary of evaluations using performance metrics for binary classification tasks that include AUC, F1, Precision, Recall, and Accuracy scores. Since all three models produced similar results, we set up a scoring system to compare each model's overall forecasting performance. 3 points indicate that the model performs the best at a particular metric, models are ranked according to their accumulated points.

From the results, XGBoost ranks the highest in terms of overall predictive performance, with logistic regression coming second, while KNN performed the worst out of all three models. Observations from Table 5 suggest that XGBoost didn't perform poorly on all metrics. XGBoost has the highest F1 score, which indicates that the model exhibits the best combined Precision and Recall

Table 4: Summary of Performance Metrics After Testing.

Algorithm	AUC	F1	Precision	Recall	Accuracy
XGBoost	0.713012	0.715867	0.697842	0.734848	0.712687
Logistic Regression	0.718193	0.673913	0.587121	0.720149	0.720149
KNN	0.706272	0.686627	0.725738	0.651515	0.707090

Table 5: Ranking of Models.

Algorithm	AUC	F1	Precision	Recall	Accuracy	Accumulated Points	Rank
XGBoost	+2	+3	+2	+3	+2	12	1
Logistic Regression	+3	+1	+1	+2	+3	10	2
KNN	+1	+2	+3	+1	+1	8	3

performance. This can be explained by the implemented algorithm it uses. GBDT often produces significant results for classification problems with high-dimension datasets that include multiple complex features. Moreover, XGBoost operates on L1 and L2 regularization techniques to reduce risks of overfitting in which KNN and logistic regression are lacking. XGBoost also allows hyper-parameter tuning to optimize forecast results, improving flexibility and scalability.

While XGBoost remains to be the top-performing model, logistic regression is an excellent alternative as it doesn't involve complicated fine-tuned parameters to produce similar results. However, it is worth mentioning that it has the lowest F1 score, implying that the algorithm is biased towards majority classes in which False Positive errors are more frequently detected.

KNN performed poorly on AUC, Recall and Accuracy metrics while producing significant F1 and Precision results. We suggest that this algorithm is not recommended as it is prone to high-dimension datasets, often resulting in meaningless estimations. When necessary, dimensionality reduction techniques such as implementing Principal Component Analysis (PCA) can be performed to counter this issue (Deegalla and Bostrom, 2006).

2.6. Feature Stability Analysis

Our initial assessment using performance metrics concluded that XGBoost most accurately predicts outcomes compared to KNN and logistic regression. This was done through train-test splitting techniques for consistent predictive performance. We wish to further perform rigorous stability analysis on the feature space of the model using k-fold cross-validation technique with k as 5 (Rodriguez et al., 2010)].

Adopting the 5-fold cross-validation measure mainly refers to the partition of the dataset into 5 subsets, with one subset as the test set and the remaining 4 subsets as the training sets. Adopting this validation test improves the accuracy of evaluating the model, resulting in more robust estimates of unseen data (Wong and Yeh, 2020). The overall accuracy of the model will possibly be comprised where feature space stability is unknown if there is no such technique implemented. Table 6 states briefly how the algorithm works to evaluate XGBoost performance.

Table 7 illustrates the results after performing 5-fold cross-validation on XGBoost.

In an overfitted model, we would expect observations of large fluctuations in ac-curacies to appear. However, results evidently indicate that there is little variation of accuracy between different iterations of training and testing. Through cross-validation of 5 iterations, we can say that our model is optimized to produce reliable predictions. Another conclusion suggests that the created feature space is stable and robust regardless of sampling variation. We recommend XGBoost for its persistent forecasting performance and reliability.

Table 6: 5-Fold Cross-Validation for XGBoost Model Accuracy.

Algorithm 1 5-Fold Cross-Validation for XGBoost Model Accuracy

Require: Dataset D, Xgboost algorithm Model Xgb
Ensure: Accuracy per fold Acc_i
1: Split D into 5 equal parts: D_1, D_2, D_3, D_4, D_5
2: **for** $i \leftarrow 1$ **to 5 do**
3: $D_{\text{train}} \leftarrow D \backslash D_i$
4: $D_{\text{test}} \leftarrow D_i$
5: $M \leftarrow Xgb\,(D_{\text{train}}\,)$
6: $Acc_i \leftarrow \text{Accuracy}\,(\text{M}, D_{\text{test}}\,)$
7: **end for**
8: **return** Acc_i

Table 7: 5-Fold Cross-Validation Results for XGBoost.

Fold Number	Accuracy
1	72.4777%
2	74.6739%
3	71.7914%
4	74.9485%
5	72.1841%

3. Sensitivity Analysis

3.1. Manipulating the Distinguishing Coefficient ζ

ζ is the distinguishing coefficient that regulates the sensitivity value to be taken into account in the calculations of GRG. This coefficient was assumed to be ζ = 0.5 for initial stability of data. We wish to investigate if there's any differences to the rankings if we alter ζ.

We can see that from Table 8, changing ζ also changes the GRG statistic, but our rankings remained consistent throughout this change. This implies that our model showcases robust stability in which the values don't exhibit change when certain parameters are tuned differently. This test statistic is key to building reliable models where the features aren't sensitive to minor changes in input parameters.

Table 8: GRG Values After Alteration of Distinguishing Coefficient.

	Grey Relational Grade (GRG)					Ranking
	$\zeta = 0.50$	$\zeta = 0.40$	$\zeta = 0.45$	$\zeta = 0.55$	$\zeta = 0.60$	
x_1	0.940	0.928	0.935	0.945	0.949	1
x_3	0.916	0.898	0.908	0.922	0.928	2
x_6	0.914	0.894	0.905	0.921	0.927	3
x_5	0.913	0.893	0.904	0.920	0.926	4
x_4	0.912	0.892	0.903	0.919	0.925	5
x_2	0.890	0.869	0.880	0.899	0.909	6
x_7	0.866	0.844	0.856	0.875	0.883	7
x_8	0.707	0.669	0.689	0.723	0.737	8

4. Model Strengths and Weaknesses

4.1. Evaluating Generalizability

When we look into how generalizable our model is, we have to take note of how relatable this model is to other kinds of matches, locations and other sports. It is known that tennis courts vary in surface material, but through our model, we suggest that this factor is not significant. The same can be said for women's matches, where the features remain consistent.

One important note is that this model is only applicable to sports with similar principles to tennis. For example, racquet sports with clear serves and returns (eg., table tennis, badminton) can be modelled with promising predictions and strategies for coaches to follow too. Since there are predictions the model didn't make right, We recommend future models include detailed data such as detailed cognitive statistics and the player's physical properties during the match.

4.2. Strengths

GRA algorithm was performed to analyze and rank the 7 features that significantly impacted the results. Data-driven mathematical calculations were done instead of cherry-picking factors or choosing based on intuition, minimizing selection bias in the process. Feature selection also removes unnecessary noise and focuses on relevant variables to be modelled. This proposed approach provides better explanatory variables that can be understood intuitively. Machine learning models are capable of integrating more features in the process for more accurate predictions. This includes predictions indicating future strategic trends and game patterns.

Instead of selecting a single model for predictions, we used multiple supervised algorithms to compare and analyze which models are most suitable for certain situations. Analytics and coaches may avoid coming to single conclusions for a match, instead, it opens up a wider range of interpretations.

4.3. Weaknesses

Although our approach is flexible for new features to provide accurate short-term forecasts, it lacks predictive accuracy on long-term outcomes. This is due to the dynamic nature of tennis games where patterns evolve over time. Current extracted features may alter in terms of importance and

relevance that are influenced by external factors we have limited data on. We encourage future researchers to input various features that may represent social, economic, natural and technological trends at the time for a more throughout study.

Another limitation suggests that the model heavily relies on continuous sourcing of new match data. It is challenging for supervised algorithms to forecast outcomes with scarce historical data. Moreover, due to the complexity of this approach, real-time integration involving constant input of data will be difficult to deploy without sufficient resources.

5. Conclusion

In this paper, we proposed a machine learning approach to forecast the flow of points in tennis matches using raw data from the 2023 Wimbledon Gentlemen singles matches. By implementing feature engineering techniques using Grey Correlation Analysis, we extracted and ranked 7 features that significantly impacted a player's victory status in the match.

Furthermore, training and testing sets were separated for model building. Our analysis introduces a general comparison of 3 different supervised learning algorithms of their forecasting performances which are K-Nearest Neighbours, XGBoost and Logistic Regression. Evaluation of these algorithms is executed through common performance metrics, such as confusion matrices, ROC curves, AUC, F1, Precision, Re-call and Accuracy scores. By adopting a scoring system for each metric, XGBoost outperforms KNN and Logistic Regression in overall predictive performance.

Sensitivity analysis was conducted by manipulating the distinguishing coefficient which showed robust consistency in the ranking of extracted features. Moreover, our approach opens up more interpretation for match prediction through features that are easier to understand. However, it provides subpar accuracy for predicting out-comes over an extended period of time. We recommend future research to include more external trends that are relevant to tennis for a more conclusive study.

Acknowledgments

We thank the Consortium for Mathematics and its Applications (COMAP) for providing the dataset in their Mathematical Contest in Modeling (MCM) competition. We thank Dr. Chin Wen Cheong (Xiamen University Malaysia) for comments on this manuscript. This research was supported by Xiamen University Malaysia Research Fund (grant no. XMUMRF/2020-C6/IECE/0016).

References

ATP. Atp rankings: Pif atp rankings (singles): Atp tour: Tennis, 2024.

Vincenzo Candila and Lucio Palazzo. Neural networks and betting strategies for tennis. *Risks*, 8(3), 2020. doi: 10.3390/risks8030068.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. doi: 10.1145/2939672.2939785.

Sophie Chiang and Gyorgy Denes. Supervised learning for table tennis match prediction, 2023.

COMAP. Consortium for mathematics and its applications (comap), 2024.

Rodney X. Sturdivant David W. Hosmer Jr., Stanley Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, 2013.

Sampath Deegalla and Henrik Bostrom. Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In *2006 5th International Conference on Machine Learning and Applications (ICMLA'06)*, pages 245–250, 2006. doi: 10.1109/ICMLA.2006.43.

Mateus De Araujo Fernandes. Using soft computing techniques for prediction of winners in tennis matches. *Machine Learning Research*, 2(3):86–98, 2017. doi: 10.11648/j.mlr.20170203.12.

Juan D. Rodriguez, Aritz Perez, and Jose A. Lozano. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):569–575, 2010. doi: 10.1109/TPAMI.2009.187.

Roselina Sallehuddin, Siti Mariyam Hj. Shamsuddin, and Siti Zaiton Mohd Hashim. Application of grey relational analysis for multivariate time series. In *2008 Eighth International Conference on Intelligent Systems Design and Applications*, volume 2, pages 432–437, 2008. doi: 10.1109/ISDA.2008.181.

Ramraj Santhanam, Nishant Uzir, Sunil Raman, and Shatadeep Banerjee. Experimenting xgboost algorithm for prediction and classification of different datasets. 03 2017.

Michal Sipko and William Knottenbelt. Machine learning for the prediction of professional tennis matches. MEng computing-final year project, Imperial College London, 2, 2015.

Qinbao Song and Martin Shepperd. Predicting software project effort: A grey relational analysis based method. *Expert Systems with Applications*, 38(6):7302–7316, 2011. doi: 10.1016/j.eswa.2010.12.005.

Sascha Wilkens. Sports prediction and betting models in the machine learning age: The case of tennis. *Journal of Sports Analytics*, 7:99–117, 2021. doi: 10.3233/JSA-200463.

Tzu-Tsung Wong and Po-Yang Yeh. Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1586–1594, 2020. doi: 10.1109/TKDE.2019.2912815.

Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Ruili Wang. Efficient knn classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1774–1785, 2018. doi: 10.1109/TNNLS.2017.2673241.

H.-J. Zimmermann. *Fuzzy Set Theory—and Its Applications*. Springer Dordrecht, 2011.