Multi-Factors Analysis Using Visualizations and SHAP: Comprehensive Case Analysis of Tennis Results Forecasting

Yuan Zhang
Physical Education Department, Northwest University, Xi'an, Shaanxi, 710069, China

***Abstract*—Explainable Artificial Intelligence (XAI) enhances interpretability in data-driven models, providing valuable insights into complex decision-making processes. By ensuring transparency, XAI bridges the gap between advanced Artificial Intelligence (AI) techniques and their practical applications, fostering trust and enabling data-informed strategies. In the realm of sports analytics, XAI proves particularly significant, as it unravels the multifaceted nature of factors influencing athletic performance. This work uses a rich data analysis flow that includes descriptive, predictive, and prescriptive analysis for the tennis match outcomes. Descriptive analysis uses XAI techniques such as SHAP (SHapley Additive exPlanations) with diverse factors such as physical, geographical, surface level and skill disparities. Top players are ranked; the trend of country-wise winning is presented for the last many decades. Correlation analysis presents inter-dependence of factors. Correlation analysis presents inter-dependence of factors. Predictive analysis makes use of machine learning models, the highest overall accuracy of 80% according to the K-Nearest Neighbors classifier. Lastly, prescriptive analysis recommends specific details which can be helpful for players and coaches as well as for overall strategies planning and performance enhancement. The research underscores the significance of AI-driven insights in sports analytics, particularly for a fast-paced and strategic sport like tennis. By leveraging advanced data analytics methods, this study offers a nuanced understanding of the interplay between player attributes, match contexts, and historical trends, paving the way for enhanced performance and informed strategic planning in professional tennis.**

***Keywords*—Artificial intelligence; data analytics; machine learning; match result prediction; XAI; SHAP**

I. Introduction

In the age when technology plays a crucial role in the world, data has become a valuable commodity in any field; and sports analytics is no exception. The extension of big data in the professional sports realm has revolutionized the way performance, planning, and decision-making process, is approached. *Sports data analysis* is vital for advancing the understanding and performance of athletic activities, providing a foundation for evidence-based decision-making in sports. With the help of advanced techniques and latest analytical tools, it helps coaches, athletes, and teams to have deep insights by identifying patterns, optimizing game plan, and get better results. The analysis includes player and game statistics, game dynamics, physiological and even psychological data and thus by bringing sports science to modern computational sports science. Moreover, the latest trends in data analysis include predictive modeling, offering insights into player fatigue, injury likelihood, and team performance under various conditions and help to predict the game outcome. Sport data analysis is being carried out in all types of sports worlds wide to gain optimal results [1]. Among sports, tennis, an aerobic and somewhat complex sport, presents an excellent chance to use data science for gaining a competitive edge and prognostic the outcomes of the matches.

Tennis is one of the most popular sports globally and is played by millions with a combination of energetics skills and physical strength, thinking ability and patience as elements such as athleticism, strategy, and power. With background foundations of 19th century, tennis has transformed into a highly competitive and technically demanding game, among both individual and team formats [1]. While other team sports tend to place their strength on the performer's abilities in the context of change and variability, such as the base depends on variant of surfaces, condition of weather, and opponents [2]. Every game turn into a battle, where a player needs to put focus on all capacities to win as fast as possible, to play using both strong and smart tactics [3]. The evolving shift of the tennis game, powered the need to enhance the technological innovations and progressive metamorphosis, continues to push the boundaries of human performance, raising tennis beyond the status of sport and turning it into both human physical and intellectual excellence based on their speed, endurance, and strength, coupled with technical precision and mental ability to boost [4]. Modern tennis involves dynamic interactions between players and surfaces, where factors such as court type, weather, and player tactics highly impact the outcomes of matches. This complexity makes tennis match a captivating sport, both as a form of entertainment and as a subject of in-depth analysis.

Match results prediction is accomplished by examining key performance indicators such as rally length, spin rates, serving efficiency, player movement, shot placement, and other factors that may influence match results. As sport evolves, data-driven approaches are becoming increasingly crucial for increasing in player performance, refining strategies of coaching, and improving in match outcomes. The integration of AI and data analytics in tennis research has opened new gateway for understanding player behavior, optimizing performance, and predicting outcomes [5]. By leveraging vast datasets that include player statistics, physical factors, match outcomes, and even skills analysis, AI models provide unpredictable in-depth

*Corresponding Author

analysis into various aspects of the game. AI-based predictive analytics use machine learning (ML) models to process data to predict outcomes based on player strengths, weaknesses, and historical performance against specific opponents. Such prediction is used by the coaches in defining training and modifying match strategies. Data analytics based on AI has been used in tennis for analysis of multi-perspective and to achieve enhanced precision [6]. In existing studies, the researchers have focused on the limb movements and force generation modes of athletes [7].

To provide a clear structure for the presentation of the research, the paper is divided into five sections. Section II provides a detailed analysis of our research contribution based on objectives. The existing literature review in Section III discusses prior research and insight research gaps. The specifics of dataset, data preprocessing, and the methods used for analysis are explained in Section IV of the study. Section V contains result and discussion, rely on descriptive, predictive and perspective analyses. The study findings are discussed in this section and related to the objectives of the research. Lastly, Section VI of the paper provides a recap of major conclusions and recommendations to extend the present study and advance the knowledge in the field.

II. Objectives

In this research study, we aim to carry out a comprehensive analysis of the tennis dataset based on real world data of more than three decades. It shares details about the exploration of various factors which may influence the match outcome. For exploratory data analysis, the factors of various perspectives are considered. The features of players are considered and then the features of losers and winners are separately considered. The distribution of aces is visualized based on diverse surface areas. The pair plot of winners and losers are explored. The correlation matrices are computed for diverse types of features. The predictive analysis consists of application of various data mining algorithms for classification which include K-Nearest Neighbor (KNN) and Ridge Classifier (RC) and Label-Propagation (LP). The results are evaluated based on accuracy, precision, recall, and F-measure. Lastly prescriptive analysis presents the recommendation of various strategies. The main contributions can be summarized as follows:

- Developed a data-driven framework that coupled with ML models on domain-specific factors to anticipate tennis match outcomes, providing actionable deep insights for players, coaches, and analysts.
- Analyzed the impact of seven key factors, including player ranking, performance ranking, player characteristics, demographic, physical health, surface level, and skill level analysis on tennis match predictions.
- Comprehensive data analytics are carried out using three diverse approaches of descriptive analysis, predictive analysis and prescriptive analysis.
- Exploration data analysis of real-world data varies out using state of the art Data visualization
- Using SHAP method for interpretation of top factors which is widely used in the latest eXplainable Artificial Intelligence perspective.
- Achievement of accuracy as high as 80% to predict the outcome of the match using lazy classifier of nearest neighbor, demonstrating its effectiveness for tennis match outcomes.

III. Related Work

The use of data analytics for sports data analysis is an active research area due to its significance [8]. As one of the most powerful machine learning algorithms, it provides the highest precision and performance when computing. It is a favorite among researchers and extensively used in several fields. A vast number of works emphasize the capacity of ML in forecasting and assessing talented tennis players and performance, following series of steps from data collection to evaluation as shown in Fig. 1. Thus, Panjan [9] considered the determination of predicted results of young athlete's skills and physical measurements as one of the ML models getting high results especially in the female sportspersons' evaluation. It was a much better way to select coaches than just picking the one that has been in the industry for a long time. Siener [10] pointed out that more concepts are relevant for consideration, and excluded physical abilities and early performance measures as specific metrics cannot adequately capture prediction models. Related to this, ML has also been used in player categorization according to their performance. Filipic [11] conducted predictive analysis to classify professional tennis players into quality groups based on the ATP rates. It helps us as coaches and the players to analyze strengths and weaknesses of team and individuals. It also pointed out that there are other success factors besides performance strategies, including mental hardness, training approaches, and psychological strength [12]. Makino et al. [13] selected the ATP singles match to analyze point winners when influenced by the court and the players' style. To illustrate the ability to practice different and more creative forms of analyzing data, Almarashi et al. [14] took a different more creative approach by demonstrating the ability to predict trends of players' performance over a period. Chen and Groll [15] used decision tree algorithm, and the results showed that it yields high level of accuracy in predicting the match outcomes for both men and women's tennis as was also discovered by Ghosh et al [16] who used logistic regression. It has also been attempted to identify some sample employing unsupervised learning methods. Whiteside and Reid [17] applied k-means clustering to decide on the best locations for aces, where data points must be grouped based on their likeness. Li et al. [18] then trained convolutional neural networks (CNNs) on images to predict batting strength and angles, due to the success of the CNN for image recognition.

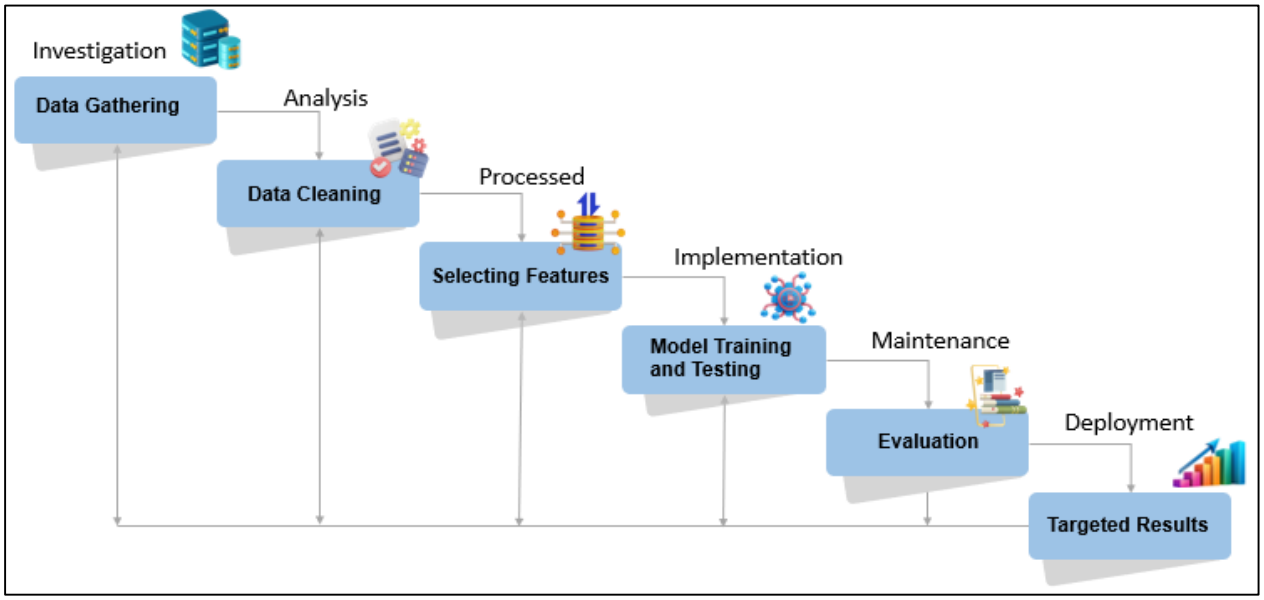

Fig. 1. An overview of the ML method, showing the iterative steps from raw data preprocessing to deploying the candidate model for applications.

In Zhou and Liu [19], different probabilities of different stances in the court were recently Bayes network predicted. Schulc et al. [20] used an LSTM network where the network learned from the video data to detect biomechanical signs of ACL injury risks. The LSTM network was able to accurately predict at-risk athletes with 75%-81% accuracy. Based on the data mining methodology, Jain et al. [21] explored sports performance, evaluated it according to benchmark models of key factors, technical aspects, and tactical difficulties confronting Chinese athletes. Together, these works establish the elaborate use of ML for the promotion of tennis proficiently in many areas such as talent recognition and changes analysis of performance and injury handicaps.

IV. Materials and Methods

In the following part of this paper, we focus on the method of this research by considering the empirical data used for collecting, cleaning, and applying for the purpose of this research, which is a prediction of tennis matches. The data used are tennis match statistical and predictive analysis, available at open-source platforms, that includes attributes based on ranking of players, their characteristics, physical factors, skills factors, surface type, tournament conditions, and match duration. In this context, the data collected is preprocessed to clean it by using data imputation methods to handle missing data and applying techniques to erase or eliminate records with many missing entries to maintain internal consistency. These methods cover three approaches for comprehensive analysis. Descriptive Analysis highlights the feature analysis. Predictive Analysis explore the correlation between parameters. By employing various models including KNN, RC, and LP to create the overall model on the factors leading to match results. KNN classifier, as working illustrated in Fig. 2, assigns a class y to a data point x, its $k-th$ nearest neighbors are determined using a distance metric $d(x, x_i) = \sqrt{\prod_{j=1}^{n}(x_j, x_{i,j})^2}$ based on predicted class neighborhood points $\hat{y} = mode\{y_i : x_i \epsilon\ Neighbors\ (x)\}$).

Furthermore, RC algorithm based on linear model that minimizes a loss function with L2 regularization $X\epsilon\mathbb{R}^{n*p}(features)$ and $y\epsilon\{-1,1\}^n$ defining predictive labels using weight vectors w corresponding to regularization strength $\alpha > 0$ for prediction $\hat{y} = sign(Xw) \rightarrow min_w||Xw - y||_2^2 + \alpha||w||_2^2$. Another graph based semi-supervised learning algorithm known as label propagation that propagates labels from labeled to unlabeled points iteratively depends on graph based-approach $G = (V, E)$ with weight matrix W and label distribution $F\epsilon\mathbb{R}^{n*c}$ based on classes $F^{(t+1)} = D^{-1}WF^{(t)}$. Prediction of labeled class computed using diagonal degree D with matrix of W. This process continues until convergence, and labels are assigned based on the maximum in F. Such models of analytics were trained and tested using historical information to provide predictions of the match outcomes with good levels of effectiveness. Furthermore, Perspective Analysis offers empirical evidence for a data-driven approach for making robust strategic planning, evaluation of performance, and decision making in professional tennis, and reveals how player characteristics and match conditions collectively determine performance.

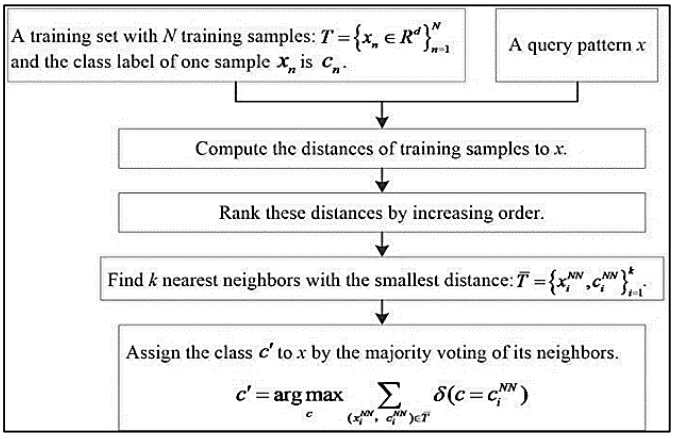

Fig. 2. The K-nearest neighbors (KNN) architecture.

V. Result Analysis

A. Descriptive Analysis

A statistical analysis of the attributes of a tennis match shows a comprehensive analysis of how several factors correlate to make an impact on a player's probability of winning and other factors as well, taxonomy shown in Fig. 3. The visualizations highlight the key attributes such as player rankings, physical fitness, and performance metrics, which collectively contribute to predicting match outcomes. All these factors have a unique function of providing an understanding of the nature of competitive tennis. The related factors of physical fitness, as shown in Fig. 4, reveal the rank ratio of winner and loser as well, highlight the ability of ranking for higher predictions. Lower ranked players relatively closer to 1 are over-represented among winners demonstrating the player ranking – which is an average of the earlier performance progress, stability and competitiveness – is perhaps the single strongest determination of match outcomes.

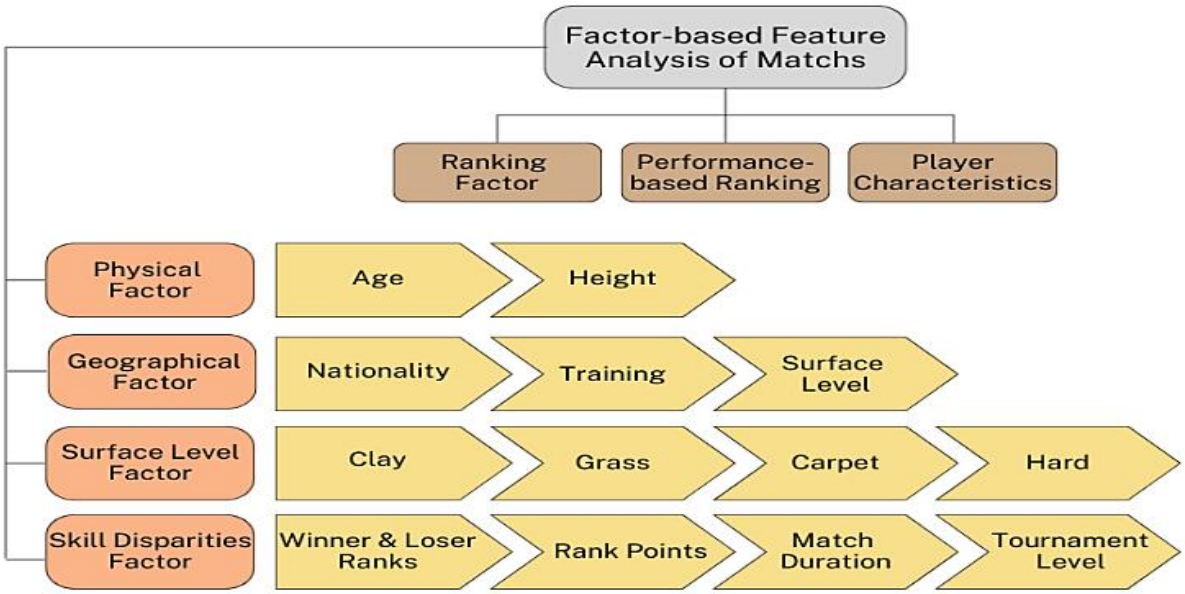

Fig. 3. Taxonomy of factor-based analysis of tennis match prediction.

This reinforces the idea that rankings are not merely statistical markers but reliable indicators of a player's performance and fitness. Similarly, features like winner and loser rank points, which are measures of ranking points acquired over a certain period, additional support sustained like ranking and competitive success as factors affecting match outcomes. Player age also emerges as a significant factor, with winners mainly under the age of 25 according to the distributions of variables winner and loser age factors. This age group characterized the peak years of physical agility, strength, and psychological resilience. This argument is further emphasized by the fact that the decline in performance observed in older players is captured by the fact that the loser age distribution tapers off, which is evidence of physicality of tennis and the fact that with age, the performance of players reduces with increasing age regardless of league ranking. Height, as captured by winner and loser height shows a more complex interaction. The distribution of player heights according to general population values, and their average of 180-190 cm indicates that height can be useful – probably in serving and court coverage – but certainly is not as definitive as ranking or age. This simply means that, though factors such as height have additional marginal utility they outweigh in their skill, strategy and mental strength.

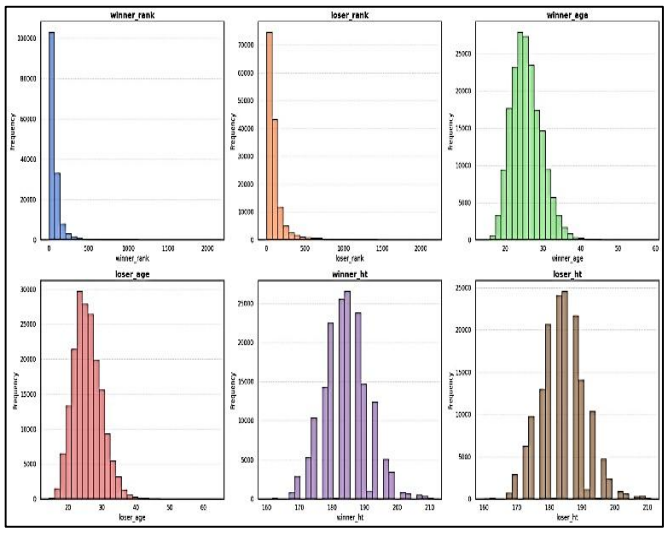

Fig. 4. Physical fitness factors affecting players performance.

The SHAP based summary plot, shown in Fig. 5, further supports these observations by showing how proposed features influence match outcomes. Factors like first, and second rank, and their associated ranking points dominate the feature

importance, emphasizing that prior performance is paramount in determining match success. Interestingly, variables such as first and second age, factors, and tournament-specific details include analysis about round, draw size, and tourney month with exhibit moderate importance meaning that even external factors affect performance, such as the tournament stage or environmental factors, can also influence progress. For instance, the level of the tournament or the number of sets played highlighted as best of attribute might benefit more experienced or physically conditioned players.

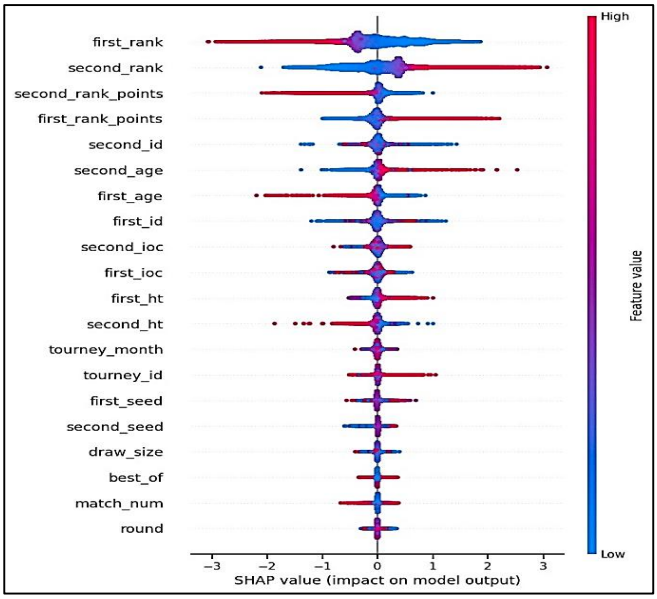

Fig. 5. SHAP analysis of features importance.

Fig. 6 provides a detailed breakdown of feature importance of predicting match outcomes. It underscores the significance of first rank pointed to winner's rank and second rank shows the progress analysis of loser's rank, guarantees that player rankings, which reflect skill, regularity, and past results, are the strongest indicators. Other features, such as second rank points highlight the ranking points of loser's players and second impactful factor first height attribute pointing to the chances of winners on the base of their height, reflecting that performance measures and physical fitness attributes are significant as secondary impact factors. The plot further underlines the contribution of other relatively significant variables with aspect of nationality background as mentioned loser's nationality code, which characteristics of the player's performance in shaping geographic or cultural patterns. The visual strength of SHAP showing how much each feature contributes to match prediction while reinforcing the idea that, although rankings overwhelm, other features bring context.

The bar chart in Fig. 7 illustrates the Top 10 Players by career wins, including legends like Jimmy Connors, Roger Federer, and Rafael Nadal, also supports these findings. These players consistently rank among the best due to their ability to maintain high performance over longer periods, which is in tune with the ranking and ranking-points identified in the evaluation of the data. The success of these players also uncovers an important part of the equation, which is psychological factor, such as mental strength and match experience, which are inferred from performance measures such as rating. The boxplot as shown in Fig. 8 depicting the distribution of aces by surface (Clay, grass, carpet and hard) highlights into how playing conditions affect serve performance. According to the distribution, Grass courts exhibit the widest distribution and median number of aces, which shows the serve on this surface is preferable for powerful players. In contrast, clay courts are characterized by a lower median and distribution, suggesting that this slow playing surface reduces the impact of aces. This information emphasizes the fact that surface type must be considered important indicating match results particularly for those players who rely on serve base. The USA dominated tennis in the last parts of the twentieth century, as was mentioned, which could also be explained by the fact that the game during that time was especially suitable for players who use powerful strikes and played fast courts, as shown in Fig. 9. But this dominance was not sustained after the year 2000, since more emerging nations approached the game with new generation of players like Spaniards and the Serbian stars who demonstrate equal powers on clay and other surfaces. Spain happened to rise steadily at the same time as its emphasis on clay court preparations, while Serbia on a similar note rose with players like Novak Djokovic. This rise of the Swiss team in the period of Federer-Wawrinka partnership show how player generations can skew national statistics. The above-presented patterns indicate that player origin and era-specific patterns are functional contextual predictors that influence match outcomes due to the general competitive conditions and training processes.

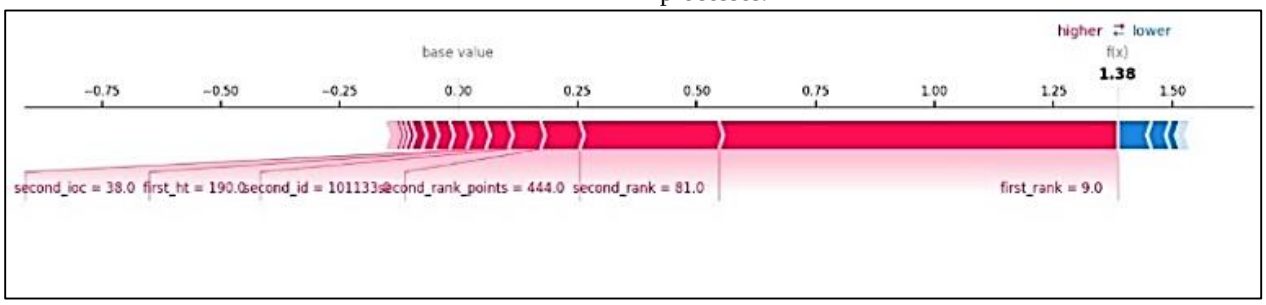

Fig. 6. Breakdown analysis of match outcomes.

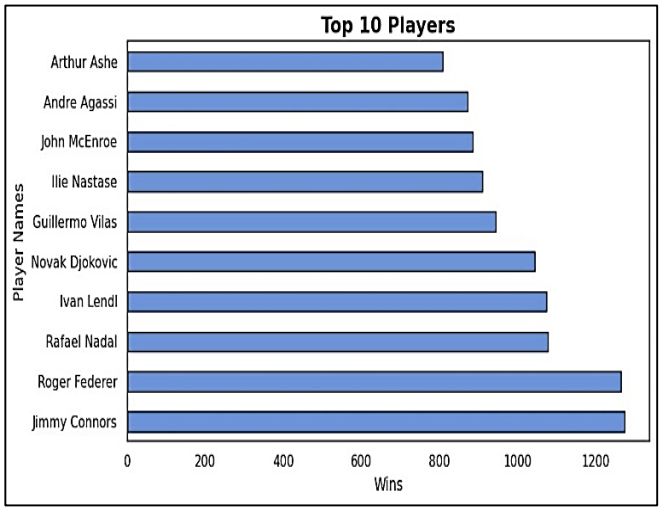

Fig. 7. Analysis of top 10 players performance.

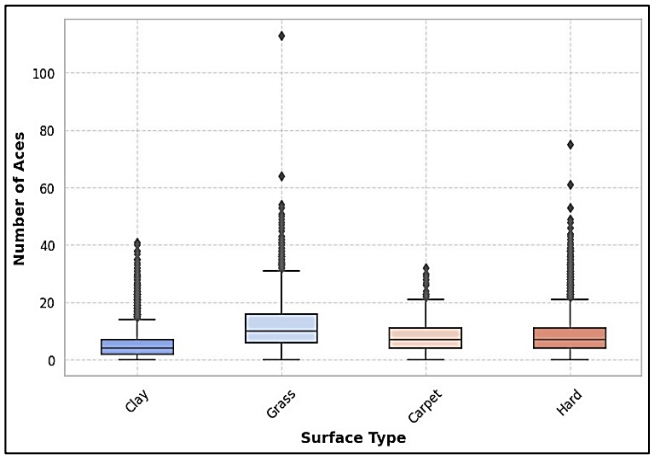

Fig. 8. Distribution of aces by surface.

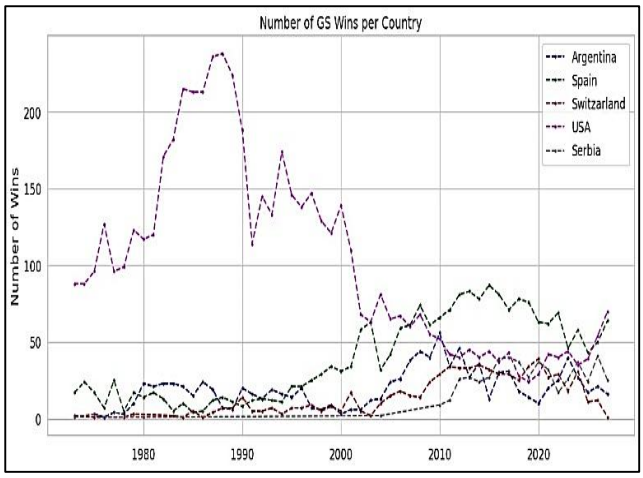

Fig. 9. Nation-wise performance of players statistics.

Analyzing variables using pair plot visualization as shown in Fig. 10, indicating factors such as winner and loser rank, match minutes that show the overall duration and tourney level indicating participating tournament score, highlighting the comprehensive analysis by examining the interplay between rankings and match intensity. Matches characterized by players with higher ranks are generally shorter, pointing to their dominance and ability to close matches efficiently. Conversely, Players with low level ranking scores tend to engage in longer matches, indicating closely contested battles where differences in skill are less pronounced. The pair plot also separates Grand Slam matches (G) from Masters tournaments (M) where the duration is usually longer because of higher level of tension, stress, anxiety, among all still showing a best ranking with physical fitness, a significant factor to ranked as winner for players. This observation illustrates that tournament setting affects matches setting, by considering the external factors other than players' characteristics predicting the matches' model.

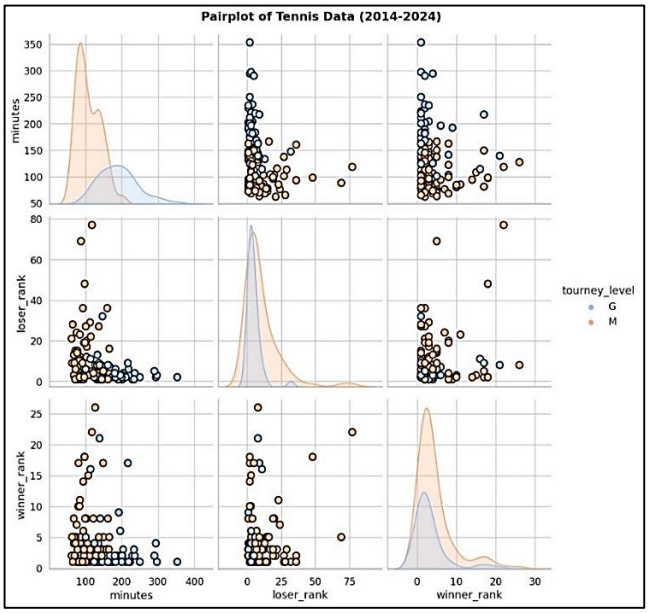

Fig. 10. Analysis of skill disparities factors.

The two correlation matrices as shown in Fig. 11 offer a comprehensive examination of the various relationships in relation to different variables in the database: tournament/match attributes and players' performance indicators. The features of this correlation include the date of the tournament; draw size; the match number; and performance indicators of the player who lost the match; aces served and committed, double faults, total serve attempts, first serves made, and break points faced. There is a very tight positive relationship between Attempts and First (0.93), suggesting that many serve points attempted deliver a high probability of successful first serve hitting. As with many of the other performance indicators, there is a strong relationship between break points saved and break points faced (0.92) – players that frequently find themselves on the wrong end of a break point usually show that they can sustain a lot of those situations. The strong positive relationship which is evident between the variable's games served and serve points attempted (r=0.94 mean, reveals the direct relationship between the number of service games played serves attempted. Low correlations between the date of the tournament and draw size and most of the performance indicators indicate that these characteristics have little influence on the result of ongoing inspired matches.

The Players Information and Performance Association Matrix looks at how certain variables in mutating with player characteristics which include the winner seed, winner height, winner rank, and match performance indicators which include aces served, double faults, break points faced, and games served by the winner, in Fig. 12. Cohesion between first serves in and serve points attempted by the winner is also evident with a coefficient value of 0.94 for the pair of variables. The finding that linkage of games served, and the break points faced by the winner (r= 0.94) suggests that the consistency of the serving players is likely to go down with the game faced on their serve as he matches progress especially in terms of break points faced. The correlation –0.33 of winner rank and winner rank points indicate that players with low numerical value of rank will tend to gain more ranking points because they have performed better than other players over the season. They both showed some relationship with match duration to some key performance indicators, and this was evident in the breakdown of the break points saved and the first serves clinched to show how stamina and service comes in handy during long drawn-out matches.

Overall study highlights the evaluating probabilities of tennis match outcomes are a complex process and factors including player rankings, age, and prior performance emerging as the most influential features for the match outcomes. Height, tournament conditions, and match dynamics provide the secondary features that add more context and depth to the predictive model bringing more of the game into the analysis. The analysis highlights that while player rankings and previous performances predicting environmental and intrinsic aspects including nature of the ground, playing duration, and players' physical characteristics including height and serve effectiveness enhance the complexity of the game adding on to the analysis. Such results suggest that more global and data-driven strategy is required to accomplish successful modeling of match results. By leveraging data analytics methods, we can build more robust systems that reflect the interplay of skill, strategy, and resilience in tennis match. This approach does not only improve the efficiency of forecast, but can also define conceptual framework for action, improvement of performance, and decision-making in professional tennis.

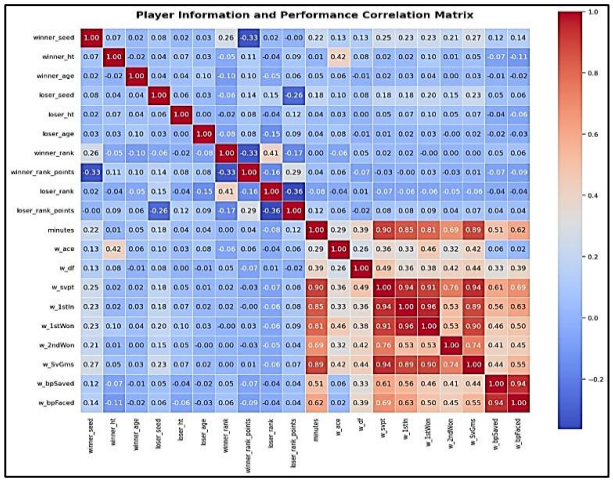

Fig. 11. Correlation analysis of player performance and tournament information.

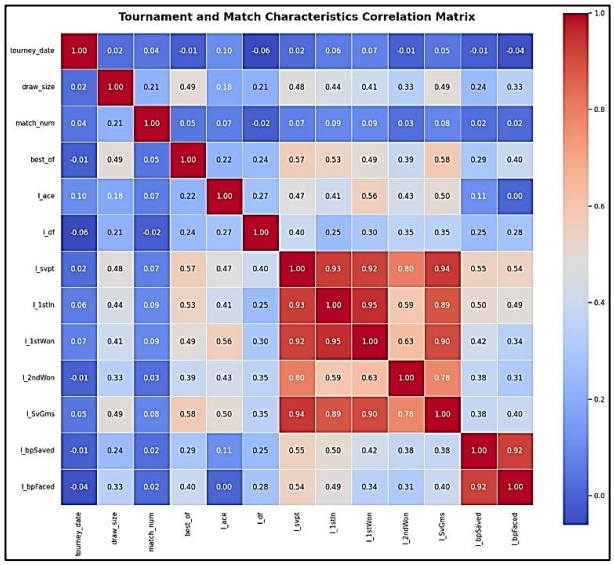

Fig. 12. The relationship between players' characteristics like age, height and ranking levels on the players' performance.

B. Predictive Analysis

The findings from the three classifiers include K-Neighbors Classifier, Ridge Classifier, and Label Propagation, in terms of accuracy, F1-score, and ROC based Area Under the Curves (AUC) values that show how the predictive models performed in the experiment with the goal to assess the strengths and limitations of using the selected algorithms for tennis match predictions. Detail analysis of results is shown in Table I.

The K-Neighbors Classifier gives the highest accuracy of 80%, with 63% F1-measure shows that the model provides a high probability of projecting the match outcome accurately most of the time, which is promising for applications where precise predictions are important. However, the values of the F1-score equal to 63% mean that the model makes moderate accurate predictions relying on players' performance with tournament levels, leads towards may some challenges in identifying less frequent match outcomes, potentially leading to some false positives or false negatives. Finally, the AUC of 78% also validates the model efficiency in determination of between the two classes of players as winner and loser but still more work is needed to be identified by the optimal decision threshold. The Ridge Classifier performs less accurate as compared to K-Neighbors Classifier achieves only 71% accuracy, utilizing both precision and recall, given its considerably higher F1-score, 69%. This implies that Ridge Classifier could be much more suitable for identifying both 'winners' and 'losers' particularly in formulation whose class distribution is skewed. Its lower AUC of 71% shows that the model does not perform as well regarding the ability to classify match outcomes throughout the probability distribution, with particular emphasis on the low success of the distinction between the positive and negative classes at various thresholds. This shows that, although the Ridge Classifier is quite balanced in terms of predicting outcomes.

Another classifier, Label Propagation tested with 77% accuracy is nearer to both models in the raw predicting power when it comes to predicting match outcomes. However, its F1-

score is 65% and it is lower than the two models we considered: K-Neighbors Classifier and Ridge Classifier which means that its precision/recall co-efficient is less accurate than these two. This has the implication that the model could be correctly classifying more instances, particularly in the minority class, resulting in either false classification as positive or as negatives. The AUC of 71% also shows us that it does not rank as high as the K-Neighbors Classifier in terms of the model's ability to show the difference between the match outcomes based on various factors, but it is better than the Ridge Classifier. Overall, Label Propagation has a reasonable level of predictive accuracy as for match outcomes, but this model has a low ability to set a moderate ratio of precision and recall as well as it has weak discrimination in contrast to other models.

TABLE I. ANALYSIS OF CLASSIFIERS FOR PREDICTIVE MATCH OUTCOMES (%)

Model	Accuracy	Precision	Recall	F1-Score	AUC
KNN	80	75	70	63	78
RC	71	72	66	69	71
LP	77	71	60	65	71

Overall, K-Neighbors Classifier surpasses the other two in its accuracy AUC, meaning that K-Nearest Neighbors Classifier, indicating that it is better at making correct predictions and distinguishing between match outcomes. However, looking at the F1-score, it can be concluded that it could be allowed better ratio of precision and recall values. The Ridge Classifier proves to improve the balance of classification but offends accuracy and AUC value. Label Propagation, while offering good accuracy again is not good in f1-measures. These results highlight the trade-offs between model performance metrics and underscore the need to select a model based on the specific requirements of the task, such as whether the priority is maximizing prediction accuracy ability to distinguish between classes, as analysis shown in Fig. 13.

C. Perspective Analysis

Let us now focus on the third type of data analytics approach of prescriptive analysis which focuses on recommending specific strategies based on data analysis. The aim of this type of analysis is to get the desired outcomes based on analysis of historical data, and application of predictive models. Unlike descriptive analysis, which explains what has happened and which is main part of this manuscript as well, and predictive analysis, which forecasts what might happen, prescriptive analysis shares the answer to the main question of what is required to be done.

Prescriptive analysis uses data-driven insights to recommend specific training and strategies tailored to players' needs and goals. For making robust strategic planning, evaluation of performance, and decision making in professional tennis, and reveals how player characteristics and match conditions collectively determine performance. By analyzing player attributes for prescriptive analytics provides actionable recommendations for optimizing performance. These insights into a player's efficiency or their success on specific surfaces can guide them to match preparation strategies. Fig. 14 shows Receiver Operating Characteristics (ROC) curve is shown for comparison and shows Area Under the Curve (AUC) too.

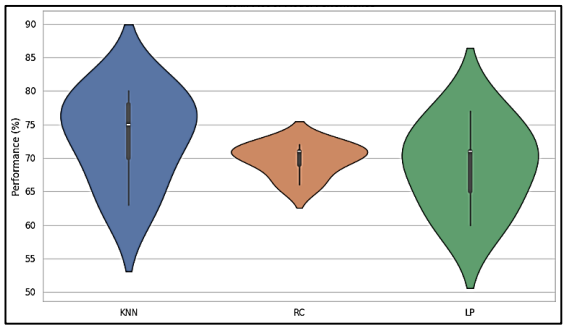

Fig. 13. This comparison reflects the accuracy measures of all applied models providing the performance evaluation of the ML approaches.

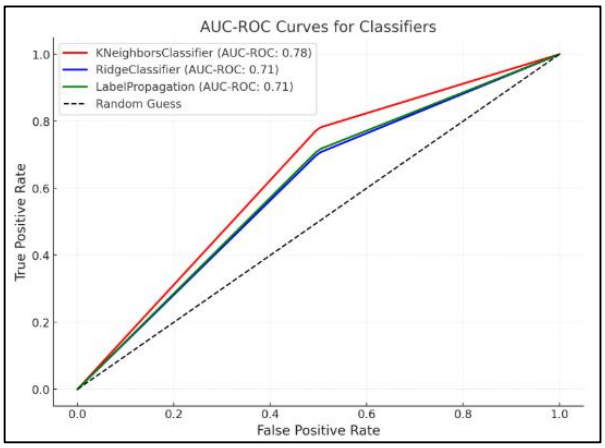

Fig. 14. The ROC-AUC curve of all applied models.

Additionally, understanding the strategies of opponent abilities and match dynamics enables players and coaches to follow strategies in real time, boost their competitive edge. Effective workload coordination in team settings to guard against injuries while maximizing on output from the players. With information concerning players, training frequency and types, courses can be constructed that would maximize their recovery period. Besides, this approach also improves the talents' performance at the personal level and proper coordination between coaches, physiotherapists, and analysts. Further, based on the same perceptions tournament organizers and stakeholders can better schedule tournaments in a way that both protects fairness of competition and players' health leading to enhanced experience. Such systems develop mutual constituencies of resources for supporting players and their sustainable performance in the sport.

In the previous works focused on tennis match result prediction by using the ML models, the numerous approaches and the features have been considered to improve accuracy, shown in Table II. The study in [22] (2022) used Logistic Regression (LR) with the features that aspect like surface type and being a winner or a loser besides the rank having an accuracy of 77%. Another approach made by [23] (2024) used Random Forest (RF) and concentrated on win/loss patterns. The suggested approach has a slightly lower accuracy of 70%.

Also, [24] (2024) put forward the Stochastic Forest Model with a player's win rate as the feature, with 74% accuracy. Work by [25] (2021) that integrated LR, DT, and RF models for changes in direction during matches gave a 75% result. Conversely, the proposed study (2024) presented the K-Nearest Neighbors (KNN) model that uses seven various features; the findings recorded a ninety percent accuracy; therefore, meaning better predictive capacity. This work sheds light on shifting paradigms of a predictive model of tennis match result based on machine learning involving feature evaluation and model selection as crucial success factors.

TABLE II. COMPARATIVE ANALYSIS WITH EXISTING STUDIES

Sr. No	Ref	Model	Features	Results (Acc: %)
1	[22] - 2022	LR	surface , Winner/Losser, Rank Rounds	77
2	[23] - 2024	RF	win/loss trends	70
	[24] - 2024	Stochastic Forest	player's win rate	74
3	[25] – 2021	LR, DT, RF	changes of direction	75
4	**Proposed – 2025**	**KNN**	**Seven Features in Fig. 3**	**80**

VI. CONCLUSION

Sports have always played a pivotal role in human culture, blending skill, strategy, and physical excellence. The coupling of artificial intelligence and data analytics into sports area highlights unparalleled opportunities to increase the power of decision-making, predict outcomes, and optimize performance rate. This research introduced three strategies of data analytics methods to investigate the factors influencing tennis match outcomes based on descriptive analysis, predictive analysis and perspective analysis, with a focus on feature-based analysis of attributes such as ranking attributes, physical attributes, and match conditions, emerges as significant predictors, providing valuable insights into the multifaceted nature of the sport. Among the models applied, the K-Neighbors Classifier achieved the highest accuracy of 80%, pointing out its potential as an effective tool for predictive analysis in tennis. This research highlights the potential of integrating advanced predictive models to help players, coaches, and analysts in strategic planning and performance optimization. Although the research study is helpful for understanding the factors for better tennis performance using XAI techniques however it is the limitation of the study that these findings are not generic and may not be applicable to other sports but only limited to tennis only. Considering the futuristic scope of the research work, let us share that the several future work can be considered for improvements can be made to increase the reliability of the predictions for more practical applications

- Higher level of data cleaning and applying diverse feature engineering techniques to refine and obtain increased quality data that would be used for developing predictive models
- Further tuning of advanced classifiers can be done by integrating hyper parameters of the classifiers to increase accuracy.
- Extending the work to conduct time-series analysis to make effective use of temporal characteristics of the data including player patterns over different tournaments.

This research not only offers understanding of the various factors of tennis match but also lays the groundwork for future explorations in sports analytical applications. This study thus clears the way for further enhancement to identify more robust methods and constructions through which data-driven approaches and models can be developed and deployed for sports and competition domains.

REFERENCES

[1] Kaur, Amandeep, Ramandeep Kaur, and Gagandeep Jagdev. "Analyzing and exploring the impact of big data analytics in sports sector." SN Computer Science 2, no. 3 (2021): 184.

[2] Liu, Sheng, Chenxi Wu, Shurong Xiao, Yaxi Liu, and Yingdong Song. "Optimizing young tennis players' development: Exploring the impact of emerging technologies on training effectiveness and technical skills acquisition." Plos one 19, no. 8 (2024): e0307882.

[3] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, Electron. Mark. 31 (3) (2021) 685–695.

[4] C. Shorten, T.M. Khoshgoftaar, B. Furht, Deep Learning applications for COVID-19, J. Big Data 8 (1) (2021) 1–54.

[5] Y. Fang, B. Luo, T. Zhao, D. He, B. Jiang, Q. Liu, ST-SIGMA:spatio-temporal semantics and interaction graph aggregation for multi-agent perception and trajectory forecasting, CAAI. Trans. Intell. Technol. 7 (4) (2022) 744–757.

[6] D.G. Ranganathan, A study to find facts behind preprocessing on deep learning algorithms, J. Innov. Image Process. 3 (1) (2021) 66–74.

[7] J. Van der Laak, G. Litjens, F. Ciompi, Deep learning in histopathology: the path to the clinic, Nat. Med. 27 (5) (2021) 775–784.

[8] Sampaio, Tatiana, João P. Oliveira, Daniel A. Marinho, Henrique P. Neiva, and Jorge E. Morais. "Applications of Machine Learning to Optimize Tennis Performance: A Systematic Review." Applied Sciences 14, no. 13 (2024): 5517.

[9] Panjan, A.; Šarabon, N.; Filipˇciˇc, A. Prediction of the Successfulness of Tennis Players with Machine Learning Methods. Kinesiology 2010, 42, 98–106.

[10] Siener, M.; Faber, I.; Hohmann, A. Prognostic Validity of Statistical Prediction Methods Used for Talent Identification in Youth Tennis Players Based on Motor Abilities. Appl. Sci. 2021, 11, 7051.

[11] Filipcic, A.; Panjan, A.; Sarabon, N. Classification of Top Male Tennis Players. Int. J. Comput. Sci. Sport 2014, 13, 36–42.

[12] Bozdˇech, M.; Zhánˇel, J. Analyzing Game Statistics and Career Trajectories of Female Elite Junior Tennis Players: A Machine Learning Approach. PLoS ONE 2023, 18, e0295075.

[13] Makino, M.; Odaka, T.; Kuroiwa, J.; Suwa, I.; Shirai, H. Feature Selection to Win the Point of ATP Tennis Players Using Rally Information. Int. J. Comput. Sci. Sport 2020, 19, 37–50.

[14] Almarashi, A.M.; Daniyal, M.; Jamal, F. A Novel Comparative Study of NNAR Approach with Linear Stochastic Time Series Models in Predicting Tennis Player's Performance. Bmc Sports Sci. Med. Rehabil. 2024, 16, 28.

[15] Dindorf, C.; Bartaguiz, E.; Gassmann, F.; Fröhlich, M. Conceptual Structure and Current Trends in Artificial Intelligence, Machine Learning, and Deep Learning Research in Sports: A Bibliometric Review. Int. J. Environ. Res. Public Health 2022, 20, 173

[16] Ghosh, S.; Sadhu, S.; Biswas, S.; Sarkar, D.; Sarkar, P.P. A Comparison between Different Classifiers for Tennis Match Result Prediction. Malays. J. Comput. Sci. 2019, 32, 97–111.

[17] Whiteside, D.; Reid, M. Spatial Characteristics of Professional Tennis Serves with Implications for Serving Aces: A Machine Learning Approach. J. Sports Sci. 2017, 35, 648–654.

[18] Li, J.; Zhang, X.; Yang, G. The Biomechanical Analysis on the Tennis Batting Angle Selection Under Deep Learning. IEEE Access 2023, 11, 97758–97768.

[19] Zhou, J.Q.; Liu, Y. Probability Prediction of Groundstroke Stances among Male Professional Tennis Players Using a TreeAugmented Bayesian Network. Int. J. Perform. Anal. Sport 2024, 1, 13.

[20] Schulc, A.; Leite, C.B.G.; Csákvári, M.; Lattermann, L.; Zgoda, M.F.; Farina, E.M.; Lattermann, C.; T˝osér, Z.; Merkely, G. Identifying Anterior Cruciate Ligament Injuries through Automated Video Analysis of In-Game Motion Patterns. Orthop. J. Sports Med. 2024, 12, 23259671231221579.

[21] Jain, Praphula Kumar, Waris Quamer, and Rajendra Pamula. "Sports result prediction using data mining techniques in comparison with base line model." Opsearch 58, no. 1 (2021): 54-70.

[22] Solanki, Shivans, Vikas Jakir, Akshay Jatav, and Dishant Sharma. "Prediction of tennis match using machine learning." International Journal of Progressive Research In Engineering Management And Science (IJPREMS) 2, no. 06 (2022).

[23] Hu, Jinming, Xiaohua Yang, Zixuan Huang, and Jinqi Xie. "Machine Learning in Tennis Match Analysis: Predicting Score Point Victor and Momentum Shift." In 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA), pp. 21-25. IEEE, 2024.

[24] Lv, Yinghui. "Research on Tennis Match Strategies Based on Machine Learning and Markov Chain Modeling." Highlights in Science, Engineering and Technology 92 (2024): 459-466.

[25] Giles, Brandon, Peter Peeling, Stephanie Kovalchik, and Machar Reid. "Differentiating movement styles in professional tennis: A machine learning and hierarchical clustering approach." European Journal of Sport Science 23, no. 1 (2023): 44-53.