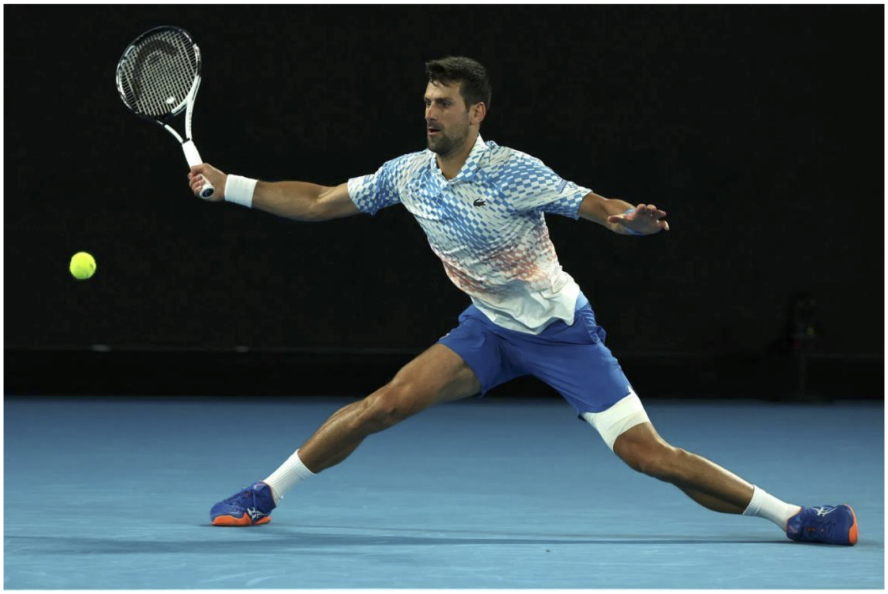
SPORTS RESEARCH

Machine Learning Based Playstyle Classification for NCAA Tennis Players



Qingyang Hou, Junying Li, Yuxing Ji, Shouzhi Wang, Ruizhe Cheng, Manqing Zhu, Xinyang Wang, Zhiyun Xu, Emma Teng

FIT x Bruin Sport Analytics

Machine Learning Based Playstyle Classification for NCAA Tennis Players

Qingyang Hou[1,2,3], Junying Li[1], Yuxing Ji[1,2], Shouzhi Wang[1,2], Ruizhe Cheng[1,2], Manqing Zhu[1,2], Xinyang Wang[1,2], Zhiyun Xu[1,2], and Emma Teng[1,2]

[1]University of California, Los Angeles
[2]Finance and Investment Technology at UCLA
[3]Bruin Sports Analytics, Tennis Consulting

1 Abstract

We present a machine learning framework for classifying NCAA tennis player playstyles using match data. Due to limited school-level data, models were trained on a professional dataset and applied to UCLA players through transfer learning. Gaussian Mixture Models generated soft labels for four playstyles, and a Random Forest Regressor achieved the best prediction performance. Playstyle scores were scaled using Universal Tennis Rating (UTR) for consistency across levels. Results were visualized via radar plots. This approach offers a scalable method for player analysis and highlights future improvements in data collection, feature expansion, and playstyle definitions.

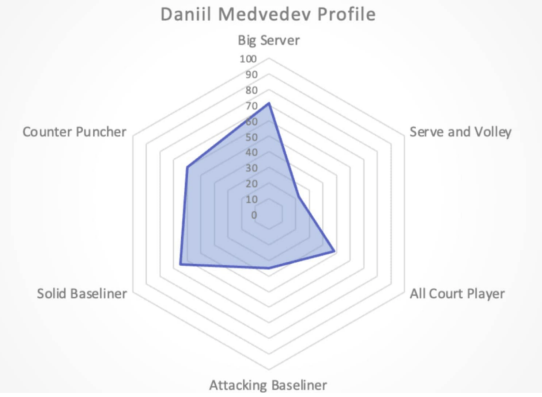
2 Introduction

We wanted to analyze the playstyles of UCLA school tennis players. Playstyle analysis can provide players with a higher-level overview of their performance in matches. The Bruin Sports Analytics tennis consulting team has been collecting player data for quite some time. We collaborated with them and used their data for school players. However, the school players' dataset was not sufficient to train a classification model. Therefore, we trained the model on professional player data found online. We then applied the trained model to school players' data to generate their playstyle analysis in a hexagon plot.

3 Playstyle Categories

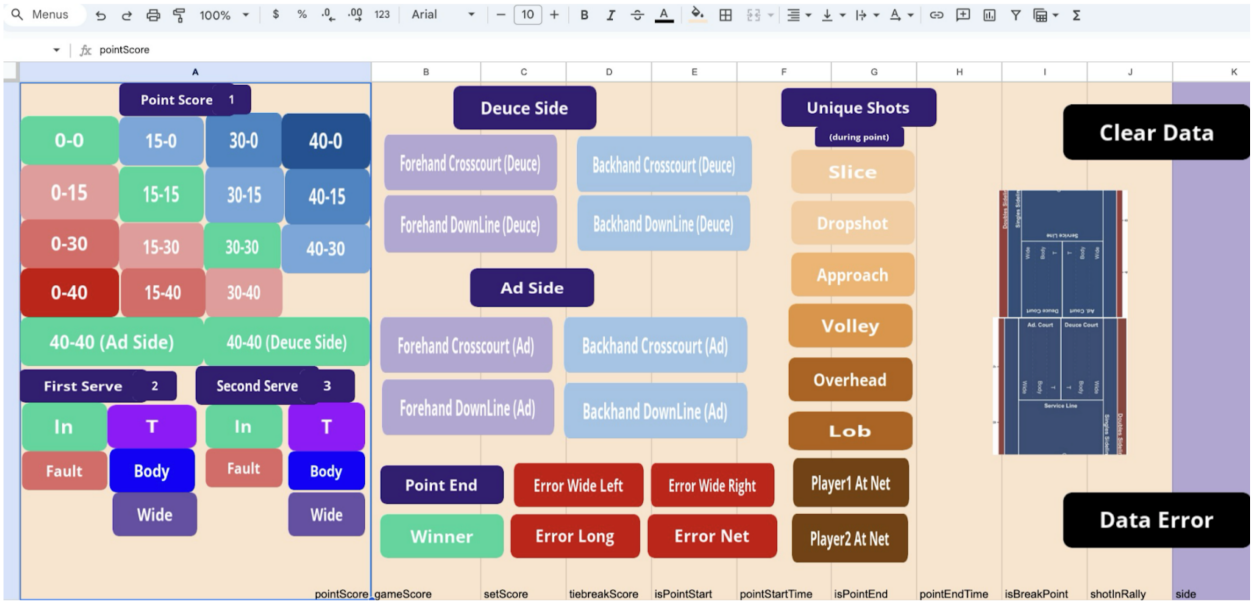
The following six categories are provided by ATP Tour. We aimed to give a score for each category for a single player. A playstyle classification example given by ATP Tour in the picture below shows the six scores for player Daniil Medvedev in a hexagon plot.

- **The Big Server:** A player with a fast first serve, who will often win points within their first two shots (e.g. aces, unreturned serves, serve + one winners).
- **Serve and Volleyer:** A player who uses serve and volley as their primary tactic.
- **All-Court Player:** A player who is comfortable in all areas of the court and often utilizes their ability at the net to their advantage.
- **Attacking Baseliner:** A player who looks to dictate play from the baseline.
- **Solid Baseliner:** A player who balances attacking and defending from the baseline.
- **Counter Puncher:** A player who is comfortable playing in defense. They use this ability to frustrate their opponent or choose their moment to turn defense into attack.



4 Data Collection

Tennis Consulting Team employed a JavaScript-based Excel interface to record point-wise match data of school players. Observers input data for each point, serve, shot, and error as they occur, facilitating detailed tracking of match progress and individual player performance. The raw data collected in Excel was further exported to CSV files. Each CSV file recorded the data for a single match.



Key Data Categories

- **Points**
  Observers initiated the recording of a point by selecting the "Point Score" button, which signaled the start of the point (`isPointStart = 1`). The point remained associated with the same `pointScore` value until the "Point End" button was selected. Points were displayed using a structured, color-coded format that represented score progression (e.g., 15-0, 30-15, deuce). The recorded scores were stored in the `pointScore` column.

- **Serves**

  - **Status Tracking**
    Observers utilized a color-coded input system to document whether a serve was "In" or a "Fault." If the serve was successful, the `firstServeIn` or `secondServeIn` variable was assigned a value of

1; otherwise, it was assigned 0. The same process was repeated for the second serve in the event of a first-serve failure. If both serves fail, the observer concluded the point by selecting "Point End."

  - **Position Tracking**
    Serve direction was recorded under `firstServeZone` or `secondServeZone` with designations of "T" (T-line), "Body" (center mass), or "Wide" (outside the service box). This information was used to evaluate serve placement strategies and their impact on match dynamics.

- **Shot Direction and Type**
  Observers categorized each shot using eight designated input buttons, organized according to the Deuce Side and Ad Side of the court. These buttons integrated parameters for shot type (Forehand/Backhand) and trajectory (Downline/Crosscourt). The system automatically tracked shot direction and type under the `shotDirection` and `shotType` columns. As the rally progressed, each shot was categorized based on these attributes, creating a comprehensive movement log of ball exchanges.

- **Errors**
  Observers classified errors based on type, selecting from predefined categories: "Wide Left," "Wide Right," "Long," or "Net." Corresponding variables (`isErrorWideL`, `isErrorWideR`, etc.) were assigned to each recorded error, enabling the identification of common error patterns.

- **Unique Shots**
  Special shot types, including Slice, Dropshot, Approach, Volley, Overhead, and Lob, were documented using binary indicators (`isSlice`, `isDropshot`, etc.). These variables helped in analyzing player shot selection and tactical decision-making during rallies.

5 Exploratory Data Analysis

Since the point-wise raw data in CSV files could not be directly used for analysis, we used Python to summarize key information into "output variables". For example, "first serve in rate", "number of first serve plus one", and "percentage of volley". These variables could be further used to calculate more output variables. Appendix A details the most relevant output variables, including their definitions and their role in classifying different playing styles.

6 The Big Server

A Big Server in tennis is characterized by a dominant and effective serve that consistently generates aces, unreturned serves, and successful serve-plus-one sequences—where the player wins the point on their first shot following the serve. This style provides a significant advantage by ensuring a strong hold on service games and exerting pressure on opponents.

The dataset includes key serve-related variables, such as `firstServeIn` (first serve success rate), `secondServeIn` (second serve success rate), `isAce` (ace occurrence), `isError` (serve errors including wide, net, and long), `serverName` (player identifier), `shotInRally` (shot sequence position), `isWinner` (winning shot indicator), and specific shot types (`isVolley, isOverhead, isApproach, isSlice`). These features collectively describe the effectiveness of a player's serve and overall rally progression.

The Serve Quality Score, a primary output variable, is derived from these input features. It integrates first and second serve statistics, including aces, unreturned serves, and serve-plus-one points, to produce a comprehensive metric representing serving efficiency.

According to ATP.com, the Serve Quality Score is calculated using two key components: the First Serve Quality Score and the Second Serve Quality Score [1].

The overall Serve Quality Score is a weighted combination of these values:

`Serve Quality Score = (First Serve Quality Score x First Serve Percentage) + (Second Serve Quality Score × Second Serve Percentage)`

7 Server and Volleyer

A Server and Volleyer employs an aggressive approach, following up their serve by advancing to the net to execute a volley with the goal of concluding the point swiftly. A volley is a shot executed near the net before the ball bounces, contrasting with a groundstroke where the ball is played after bouncing. This playstyle is designed to disrupt opponents early in the rally and secure quick points.

The Serve-and-Volley Score is used to classify players who specialize in this style. It is calculated based on two key variables: the percentage of serves followed by a net approach (`per_serve_plusOne_netShot`) and the percentage of points won through this tactic (`per_win_serve_plusOne_netShot`). These variables indicate both the frequency of serve-and-volley plays and their success rate.

The Serve-and-Volley Score is computed as:
`Serve-and-Volley Score =`
`(per_serve_plusOne_netShot x 0.5) + (per_win_serve_plusOne_netShot × 0.5)`

8 Challenges in Categorizing Additional Playstyles

The remaining four playstyle categories—All-Court Player, Attacking Baseliner, Solid Baseliner, and Counter Puncher—are more complex to quantify and cannot be determined using simple percentage-based metrics. Unlike the Big Server and Serve-and-Volley playstyles, which can be measured by distinct serve and net-play frequencies, these styles rely on a broader combination of tactics, strategies, and shot selections.

The difficulty in distinguishing between these styles arises from overlapping statistical characteristics. For instance, Aggressive Baseliners and Solid Baseliners both exhibit high groundstroke frequencies, making differentiation challenging. Additionally, All-Court Players incorporate multiple elements from various styles, requiring a flexible classification approach. Given these complexities, machine learning techniques, particularly regression models, were applied to predict playstyle scores based on input features.

9 Limitations in Data Utilization

1. **Insufficient Data for Model Training** One major limitation is the relatively small dataset available for school-level players. Machine learning models typically require extensive datasets to identify complex patterns. Since school-level datasets include only a limited number of players and matches, data scarcity may reduce model accuracy and generalizability. To address this issue, we incorporated datasets of professional tennis players who are on ATP tour from The Match Charting Project [2]. These datasets, comprising over 350 professional players' match statistics, offer a more comprehensive foundation for training machine learning models. School players are competing in NCAA Division I, which is a similar level to professional tennis. By employing transfer learning techniques, we adapted the professional-level model for school-level player predictions.

2. **Lack of Labeled Data** Another challenge is the absence of predefined labels for the four additional playstyles in the professional player dataset. Since the regression model operates as a supervised learning algorithm, labeled data is essential for effective training. To overcome this limitation, we applied Gaussian Mixture Model (GMM) clustering to identify distinct player styles based on their statistical profiles.

10 Gaussian Mixture Model (GMM) Approach

GMM is a probabilistic clustering method that models the professional player dataset as a mixture of Gaussian distributions, each representing a different player category. A 'soft' GMM variant with regularization and tied covariance constraints was used to improve clustering robustness. The model assigned probability scores to each player for belonging to a particular category. By multiplying these probabilities by 100, we derived final playstyle scores, which were subsequently used to generate a hexagonal playstyle visualization. Results of 10 random professional players are shown in Table 1.

Player	Counter Puncher	Attacking Baseliner	All-Court Player	Solid Baseliner
Marin Cilic	8.10e-05	9.999929e+01	6.23e-04	7.00e-06
Brian Teacher	6.20e-17	1.390027e-09	1.09e-19	100.000000
Marcel Granollers	3.34e-04	9.999637e+01	2.97e-04	3.00e-03
Lleyton Hewitt	3.79e-06	9.998140e+01	1.86e-02	1.20e-05
Damir Dzumhur	1.29e-04	9.970602e+01	2.94e-01	1.48e-04
Julien Benneteau	4.35e-06	9.999955e+01	3.97e-04	5.10e-05
Alexander Bublik	2.35e-08	9.998388e+01	4.19e-04	1.57e-02
Dan Goldie	4.01e-13	3.132428e-06	3.05e-14	99.999997
Ivan Ljubicic	1.67e-06	9.982983e+01	1.70e-01	6.40e-05
Jannik Sinner	7.34e-05	9.999819e+01	1.70e-03	3.00e-05

Table 1: Player classification percentages across different play styles

Despite obtaining numerical scores for all four categories, category labels remained 'unidentified. To resolve this, one method we tried was the 'hard' GMM method, which assigns each player to a single category. The ten highest-scoring players per category were analyzed to determine the defining statistical features of each group. These features were then matched with known playstyle attributes to establish category labels. However, we encountered several problems, such as no information or recorded matches found for some players, same players appear in more than one category. Then, we tried to determine the category labels using statistical analysis.

11 Classification of Player Categories

We plotted a bar plot containing average values of several output variables from each category, as shown in Figure 1

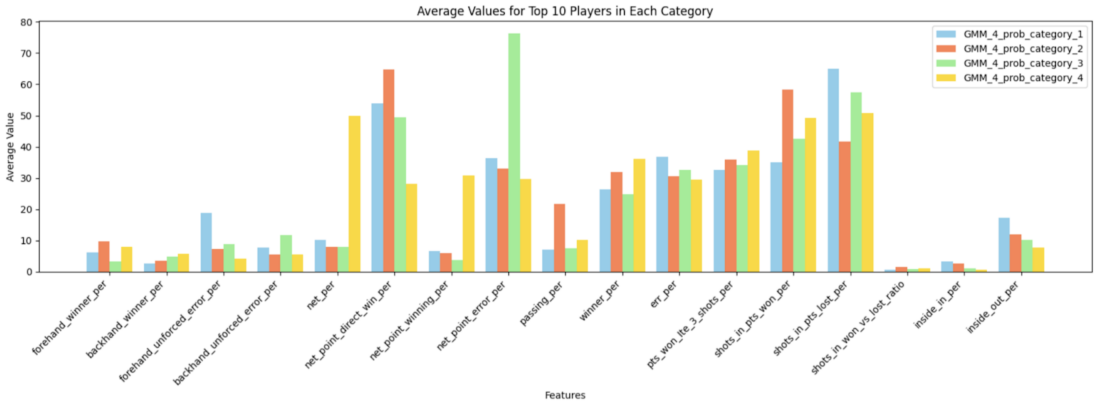


Figure 1: Bar Plot of Average Values of Output Variables

Based on the bar plot, the four GMM-generated player categories were mapped to established playstyles as follows:

- **All-Court Player:** Strong balance between baseline and net play, high winner percentages, frequent net approaches, and effectiveness in short rallies.
- **Attacking Baseliner:** Dominates points through aggressive baseline strokes, high winner rates, and preference for short rallies.
- **Solid Baseliner:** Maintains consistency with minimal unforced errors, excels in extended rallies, and balances offense and defense.
- **Counter Puncher:** Adopts a defensive strategy, forcing opponent errors while excelling in passing shots and long rallies.

12 Random Forest Regressor for Playstyle Prediction

Once playstyle labels were determined, a Random Forest Regressor was trained and tested on the labelled professional player dataset. The model achieved:

- **Training Set**: Mean Squared Error (MSE) = 107.9620, Mean Absolute Error (MAE) = 3.7131
- **Test Set**: Mean Squared Error (MSE) = 155.7973, Mean Absolute Error (MAE) = 3.9527

Alternative models, including Ridge Regression and Feedforward Neural Network, were tested but exhibited lower predictive accuracy. Results are shown in Table 2.

Model	**rMSE**	**R^2 Score**
Ridge Regression	18.17	0.432
Feedforward Neural Network	16.01	0.455
Random Forest Regressor	12.48	0.542

Table 2: Model performance comparison based on rMSE and R^2 Score

Random Forest outperformed the other models likely because the relationship between player features (e.g., shot stats, movement, consistency) and playstyle scores is highly nonlinear and feature-interactive. Unlike Ridge Regression, which assumes linearity, and the neural network, which requires extensive tuning and large data to generalize well, Random Forest is better suited for capturing complex patterns in smaller or moderately sized datasets like ours. Its ensemble structure also helps handle noisy features and avoid overfitting, making it well-matched for the multi-label regression task in our playstyle prediction project.

13 Final Model Application and Universal Tennis Rating (UTR) Adjustment

To align the performance scores of UCLA tennis players with those of professional ATP players, we employed the Universal Tennis Rating (UTR) as a scaling factor. The UTR provides a standardized measure of player skill on a 1.00 to 16.50 scale, facilitating direct comparisons across different levels of play.

In our analysis, the average UTR for ATP players was 15.51, while UCLA players averaged 13.01. To adjust for this disparity, we calculated a scaling factor:

$$\begin{aligned} \text{scaler} &= \text{average\_atp\_utr}/\text{average\_ucla\_utr} \\ &= 15.51/13.01 \\ &\approx 1.192 \end{aligned}$$

Applying this scaler to the predicted playstyle scores of UCLA players ensures that their performance metrics are proportionally adjusted to reflect the higher competitive standards of professional players. For instance, if a UCLA player's initial "Attacking Baseliner" score was 80, the scaled score would be:

$$\begin{aligned} \text{scaled\_score} &= 80 \times 1.192 \\ &\approx 95.36 \end{aligned}$$

This adjustment enhances the comparability of performance analyses between NCAA and ATP athletes.

14 Examples of Applications

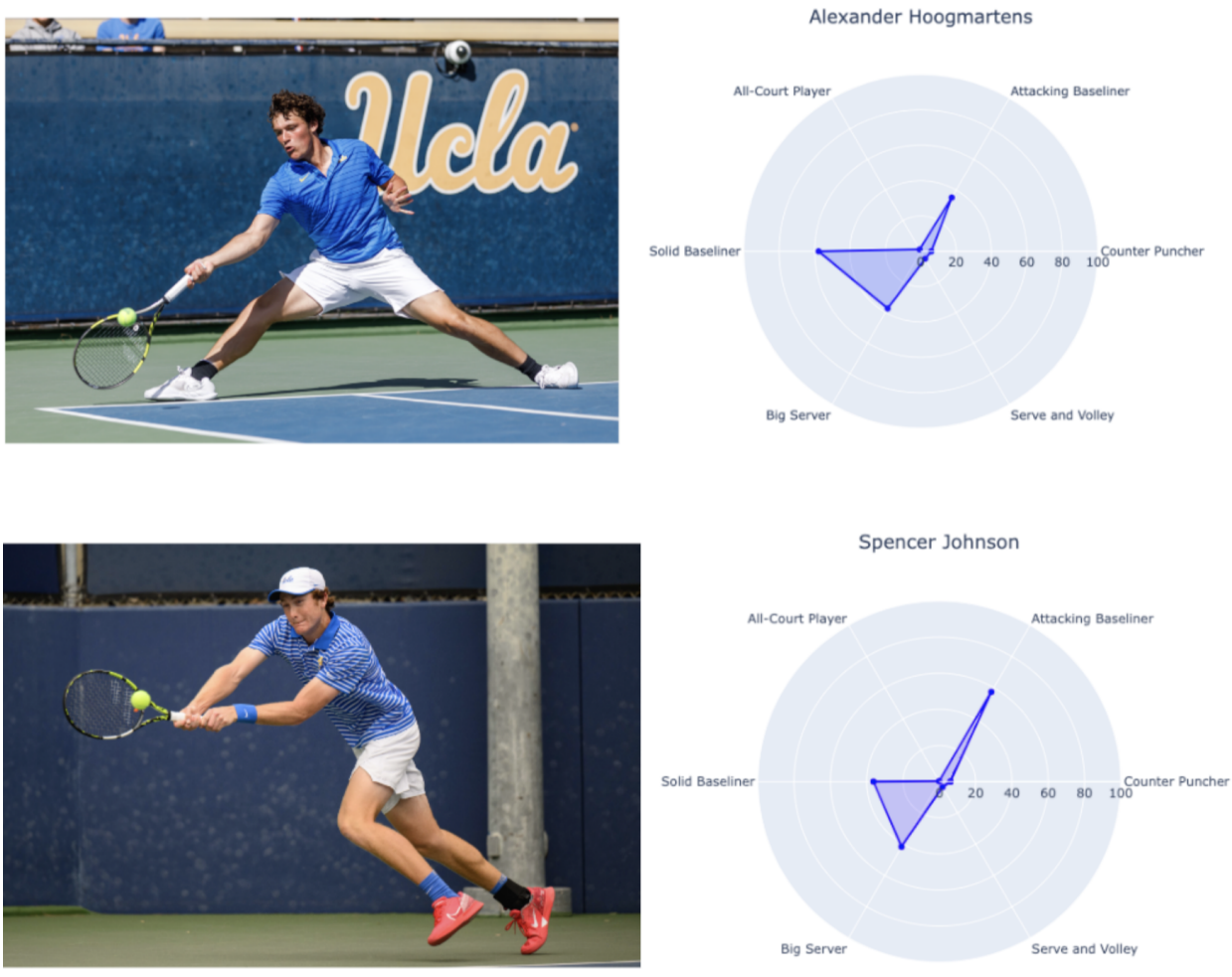


Figure 2: Bar Plot of Average Values of Output Variables

15 Discussion

While our model successfully predicts player playstyles using NCAA and ATP data, there are key areas for improvement. First, the manual data collection process is time-consuming and error-prone. Automating it with computer vision tools would improve accuracy and scalability. Second, the professional dataset lacks certain features that could enhance prediction accuracy. Incorporating more detailed metrics like player

movement or spin rate would improve the model's performance. Lastly, the current playstyle categories show overlap, particularly between baseline-based styles. Defining new or hybrid categories could reduce ambiguity and better capture diverse player behaviors. Future work should address these limitations to improve precision and practical application.

16 References

[1] *Insights: Serve effectiveness: ATP tour: Tennis.* ATP Tour. (2023, October 5). https://www.atptour.com/en/news/insights-serve-effectiveness

[2] Jeff Sackmann. (n.d.). *The Match Charting Project.* GitHub. https://github.com/JeffSackmann/tennis_MatchChartingProject?tab=readme-ov-file

17 Appendix

Table 3: Descriptions and play styles associated with key output variables in tennis analytics

Output Variable	Processing Explanation	Methods Used In
shotsPerWinningPoint_avg	Calculate the average number of shots taken by the player in every point they win	Counter Puncher
Serve_unreturned	Total number of unreturned serves; serve is not an ace, opponent touched it but did not return	Big Server
Serve_plusOne	Total number of first–serve plus one points (serve wins point with one additional shot)	Big Server
return_win_point	Total number of returns that are winners or force the opponent into an error	Aggressive Baseliner
serve_and_volley_per	Percentage of serve-and-volley plays among all serves	Server and Volleyer
serve_and_volley_success_per	Percentage of points won by serve and volley	Server and Volleyer
serve_and_volley_score	Final score evaluating the player as a server and volleyer	Server and Volleyer
per_Serve_plusOne_netShot	Percentage of serve plus one net shots among all shots	Server and Volleyer
per_win_plusOne_netShot	Total number of points won by serve plus one shot at net	Server and Volleyer
net_shot_percentage	Percentage of net shots	Server and Volleyer
Passing_per	Percentage of passing shots (baseline passes net player; winner or forced error)	Counter Puncher, Solid Baseliner
groundstrokeIn_per	Percentage of forehands and backhands that go in	Solid Baseliner, Aggressive Baseliner
direction_change_percentage	Percentage of shot direction changes	All Court Player, Aggressive Baseliner, Counter Puncher
gini_shot_selection	Gini coefficient of shot distribution (0 = equal, 1 = unequal)	All Court Player
isError	isError = 1 if any: IsErrorLong, IsErrorWideR, IsErrorWideL, IsErrorNet	Aggressive Baseliner
error_per	Total error percentage	Aggressive Baseliner
baseline_per	Percentage of shots taken from baseline	Aggressive Baseliner, Counter Puncher, All Court Player
net_per	Percentage of shots played at net	Server and Volleyer, All Court Player, Big Server
uniqueShotType_winning_per	Percentage of unique shot types that result in winning points	All Court Player, Counter Puncher, Server and Volleyer
uniqueShotType_direct_win_per	Percentage of unique shot types that directly lead to points (aces, winners)	Big Server, Server and Volleyer, Aggressive Baseliner
uniqueShotType_winByError_per	Percentage of unique shot types leading to points via opponent errors	Counter Puncher, All Court Player
winner_per	Percentage of total shots that are winners	Aggressive Baseliner, Big Server, All Court Player
inside_out_forehand	Number or percentage of inside-out forehands	Aggressive Baseliner, All Court Player, Counter Puncher
down_the_line	Number or percentage of down-the-line shots	Aggressive Baseliner, All Court Player, Counter Puncher

Continued on next page

Output Variable	Processing Explanation	Methods Used In
plus_opponent_slice_per	Percentage of points won against opponent slice shots	Counter Puncher, All Court Player, Server and Volleyer
baseline_plus_net_shot_per	Percentage of points combining baseline + net shots	All Court Player, Server and Volleyer
long_point_num	Total number of long rallies (extended points) played	Counter Puncher, All Court Player
return_attack_num	Total number of points where the player attacks off the return of serve	Aggressive Baseliner, Counter Puncher, Big Server