Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2412660

Game Changers: Unraveling the Momentum Mystery in Tennis

Summary

For years, people have discussed the impact of **momentum** on players' performance in sports. A team or player may feel they have the momentum, or "strength/force" during a match. They may use the momentum to score consecutive points for a period of time or to buck the trend at a critical juncture. Therefore, maintaining a good momentum is very important for the improvement of their performance.

In this paper, we try to build models to visualize player performance on the field, prove the existence of momentum quantify it, and then identify the influencing factors that create or change momentum during the game.

For Task 1, we want to identify which player performs better at a given time in the match and how well he performs. Whether a player can win this point is determined by the player's performance and the serving effect(player serving is more likely to win), so we used the **SARIMA** model to eliminate the seasonal impact of serving and receiving in time series data and used regression fitting to obtain the actual trend of the player's performance state and visualized it.

For Task 2, in order to demonstrate the existence of momentum in a tennis match, we discussed the **correlation** between the player's current state and the previous state, which can be shown by the **autoregressive coefficients** $\beta_3, \alpha_1, \alpha_2, \alpha_3$ in the Task 1 model were significantly different from zero.

For Task 3, we used a **Random Forest regression** model to predict momentum fluctuations during the game. We then performed a **sensitivity analysis** and **Feature Importance Detection** on the model to identify three main influences: unforced errors, the impact of winning points, and holding serve.

For Task 4, we utilized several indicators to test the accuracy and generalization ability of our model with a final and a AusOpen women's single game, among which, MSE = 0.075, R-squared = 0.699, F1 Score =0.928, Recall = 0.935. we also use confusion matrix and ROC curve to visualize. We demonstrated that our model performed well and effectively solved our forecast-type problem. We gave some discussion on the future model.

Finally, we analyzed the strengths and weaknesses of the model, wrote a memo to explain to the coaches the role of momentum in sportsthey can affect the mental state and competitiveness of sports teams and individuals as well as the outcome of the game) and advised them on how to prepare players to respond to events that impact the flow of play during a tennis match.

Keywords: Momentum; SARIMA; Time Series; Random Forest Regression

Contents

1 Introduction

1.1 Background

Wimbledon, a prestigious tennis event, captivates global audiences. Carlos Alcaraz's victory over Novak Djokovic ended Djokovic's Wimbledon dominance since 2013, highlighting the influence of "momentum." The match featured dramatic shifts, emphasizing the pivotal role of momentum in sports dynamics.

Momentum, defined as "strength or force gained by motion or a series of events," plays a crucial role in sports, influencing both psychological and physiological aspects. Studies suggest that momentum can shape the dynamics of a sporting event, impacting outcomes. Teams or athletes with momentum often exhibit higher morale, maintaining a positive mindset and psychological pressure on opponents. This advantage can lead to overcoming disadvantages, reversing critical points, and shifting the game's trajectory.

Momentum also influences match strategy. Teams with momentum tend to adopt an aggressive style, taking risks and capitalizing on opportunities, catching opponents off guard. Conversely, teams facing a shift in momentum may adopt a more conservative approach, focusing on defense and minimizing risks.

Exploring the impact of momentum in sports holds profound implications. Analyzing momentum factors aids coaches in developing relevant game tactics, enhances athletes' competitive levels, and consistently influences their performance positively. This exploration extends beyond statistics, delving into human performance, psychology, and the dynamic relationships in athletics.

1.2 Restatement of the Problem

To delve deeper into the impact of momentum in sports, we are required to do the following tasks in this paper:

- Task 1: We need to synthesize and quantify the factors that influence the score of a match and develop a model to evaluate how well a player performs at a given time in a match. Also, we will provide a visualization based on our model to depict the match flow.

- Task 2: Evaluating the Role of Momentum: We use our model to demonstrate that game situation fluctuations and player scores or wins are not random but are influenced by momentum.

- Task 3: Predicting Momentum Shifts in a Match: Develop a model to predict fluctuations in match situations, find out when momentum will shift from one player to another, and find the most relevant factors. Then, help the player perform better in the following matches, based on the differences in the fluctuations in a player's momentum from past matches.

- Task 4: Test the generalizability of the model: Apply the developed model to different types of matches and evaluate its accuracy for momentum fluctuations in matches. Analyze the limitations of the model in depth and suggest improvements for future models.

- Task 5: Finish the report and write a memo to summarize the role and importance of momentum, and give coaches advice on how to incorporate momentum into developing more effective tactics and strategies, controlling emotions to maximize the momentum advantage.

1.3 Literature Review

For more than 50 years, sports researchers have been interested in the concept of momentum, and the study of the sports literature suggests that momentum is viewed either as a "hot hand," (which is "success breeds success") or as a "psychological state".[1] However, many others did not observe any significant effects related to such a phenomenon. So there is conflicting evidence as to whether momentum is real or whether it is simply a misunderstanding among players and spectators[2] Avugos et al. conducted a meta-analysis of 250 studies and concluded that current performance does not have a significant influence on future performance.[3] However, as discussed below, some researchers are nonetheless convinced that a driving force exists, and they have proposed an alternative conceptualization of momentum as a "psychological state." A study by Martin I. Jones and Chris Harwood suggests that a range of triggers initiated perceived psychological momentum and highlighted how individuals perceived psychological momentum based on actual competitive sports experiences.[4] Factors such as motivation, confidence, optimism, energy, score changes, referee decisions, and even spectator reactions can impact momentum.[5][6] A recent investigation by Ben Moss and Peter O'Donoghue suggests that momentum affects different players in different ways that have implications for coaching and psychological support for tennis players.[7] Tennis coaches and players need to understand events that may initiate positive or negative momentum and give their hypothesized importance in determining match outcomes.

1.4 Our Work

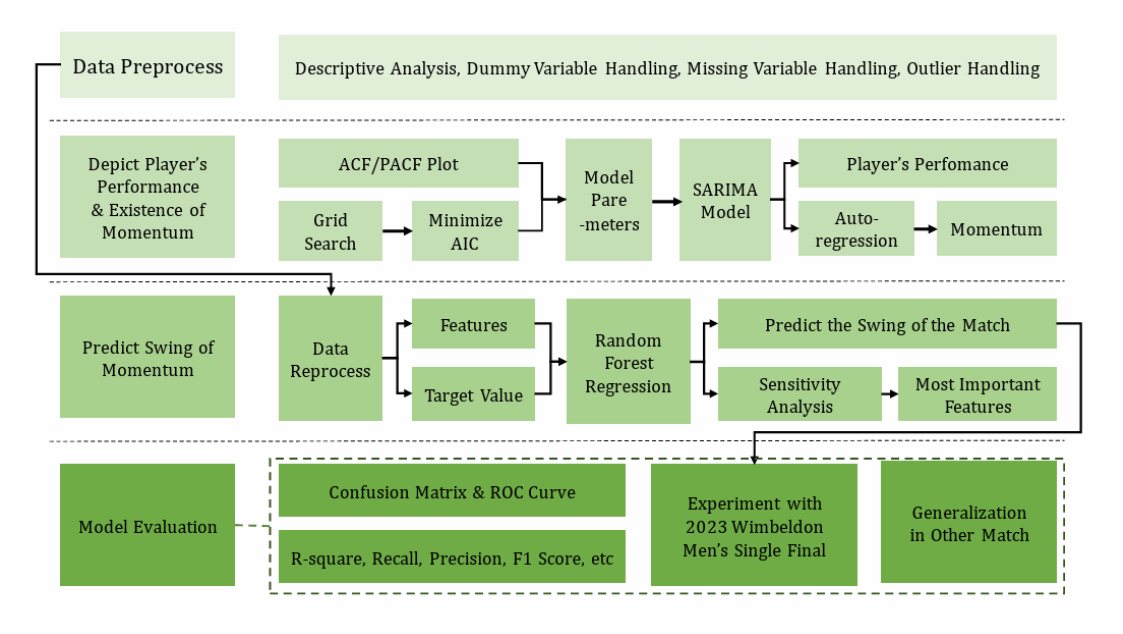


Figure 1: Flowchart of Our Work

2 Assumptions and Justification

- **Consistency in Player Performance:**
  The assumption is that the performance of players in matches is somewhat consistent. This includes factors such as players' skill levels, form, and health conditions.
  Justification: Assuming player performance consistency is reasonable, but unforeseen events like injuries or form fluctuations are possible. Historical player data can be used to adjust for such inconsistencies.

- **Correlation of Features:**
  The assumption is that the selected features are correlated to some extent with match outcomes. For example, a player's historical performance and technical indicators may be correlated with the probability of winning a match.
  Justification: Strong correlations between features and match outcomes are essential for building an effective model. For instance, past win rates and key performance indicators directly influence future match success, enhancing predictive accuracy.

- **Stability of Match Environment:**
  The assumption is that the match environment and conditions are stable to some extent unless specific factors, such as venue or weather, are explicitly considered.
  Justification: Assuming some stability in match conditions aids in constructing a simpler, more generalizable model. However, if significant fluctuations occur, the model may require more intricate handling, possibly involving interaction terms or other methods.

- **Normality of Data:**
  The assumption is that the performance status of a tennis player, along with other match-measured data, follows a normal distribution.
  Justification: Assuming a normal distribution is helpful for certain statistical inferences and model interpretations. In regression analysis, a normal distribution assumption is often used for parameter estimation and hypothesis testing. It's essential to note that the normal distribution assumption is not a strict requirement, especially in cases with large sample sizes where the model's assumptions about data distribution are more flexible.

3 Notations

Table 1: **Notations used in this paper**

Symbol	Definition
y_t	player's current scoring flow
S_t	impact of serving and receiving
P_t	player's current performance statu
$rnpt_{t,w}$	rolling net point scored
npt_t	net point scored
p	non-seasonal autoregressive order
P	seasonal autoregressive order

Symbol	Definition
d	number of non-seasonal differences
D	number of seasonal differences
q	non-seasonal moving average order
Q	seasonal moving average order
ϵ_t	a serie of white noise
$p_s w$	probability of winning a serving point
$p_r w$	probability of winning a receiving point

4 Data Preprocessing

File "Wimbledon_featured_matches.csv" contains the data set of Wimbledon 2023 Gentlemen's singles matches after the second round. Each row consists of factors such as speed of serve, number of shots during the point, direction of serve, category of untouchable shot, and so on. File "data_dictionary.csv" contains a description of the data set, providing an explanation of each column in "Wimbledon_featured_matches.csv".

First, we use `pandas` to read the data and store it in a `DataFrame` structure. Through analysis, we found a total of **7284** rows of data with **41** columns. These columns can be roughly categorized into four **4** types:

- **Text**: "match_id", "player1", "player2"
- **Numeric:** "speed_mph", "p1_distance_run", "p2_distance_run", etc.
- **Serial Number:** "set_no", "game_no", "point_no", etc.
- **Dummy Variable**: "server", "point_victor", "p1_ace", "p1_double_fault", "serve_width", etc.

Dummy Variable Handling

It is worth noting that, we need to convert dummy variables into one-hot encoding which helps make them more suitable for data analysis. Here is how we convert to one-hot encoding:

$$server = \begin{cases} 0 & \text{player1 serve} \\ 1 & \text{player1 receive} \end{cases} \quad \text{serve width} = \begin{cases} 0001 & \text{body} \\ 0010 & \text{body center} \\ 0100 & \text{body wide} \\ 1000 & \text{center} \\ 0000 & \text{wide} \end{cases}$$

Imputation of Missing Values

Next, we check if there are any missing values in the data. We found missing values in **"speed_mhp"** column, and the proportion of missing values is **0.10324**. The occurrence of

missing values is due to the player making a double fault, so the serving speed for this point was not recorded. According to the rules of tennis matches, double fault equals a bad quality serve that provides the opponent an opportunity to counterattack. Based on that, we will impute the missing values in this column with the **minimum value**.

Outlier Handling

To identify outliers, create a boxplot for numeric variables("speed_mph", "p1_distance_run", "p2_distance_run").

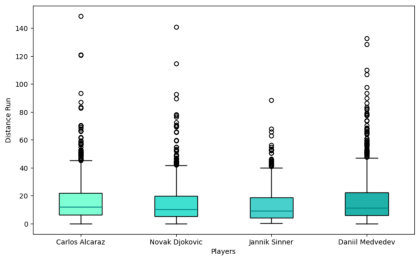


Figure 2: speed_mph

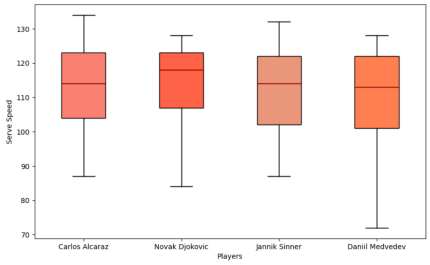


Figure 3: distance_run

Figure 4: boxplot of semifinal player

The above figures are the boxplot of semifinal players' serve speed and run distance. No outlier was detected in the 'speed_mph' figure. Outliers are present in the 'distance_run' figure. The reason for the occurrence of an outlier is that the player engaged in a prolonged rally with multiple strokes and active movement at some points. Outlier is related to tactical strategy and game situation. Thus, we shouldn't remove these outliers.

5 Task1: Depicting Player's Real-time Performance

5.1 Problem Analysis

In this task, we are required to establish a model to quantify players' performance at a given time in the match and identify which player is performing better. In particular, we should factor in that player serving has a much higher probability of winning in tennis.

The player's performance and status are directly reflected in their scoring situation. However, due to the serving effect(a player serving is more likely to win), scoring may exhibit deviations in reflecting the player's actual status. That is to say, a player losing a point does not necessarily indicate that he performs worse. Vice versa. **Thus, We need to eliminate the impact of serving efficiency on the scoring situation to obtain the player's true performance state**.

According to tennis rules, players take turns to serve. **This indicates that the impact of serving and receiving is subject to a seasonal alternation**. Therefore, we use **SARIMA (Seasonal AutoRegressive Integrated Moving Average)** model to eliminate the seasonal impact of serving and receiving. Then, regression fitting is applied to obtain the true trend of the player's performance state.

5.2 Establishment of SARIMA Model

Verifying Serving Advantage

First, we should verify our assumption that serving plays an important role in tennis matches by analyzing the given data " Wimbledon_featured_match.csv". The probability of the serving side winning the point is **0.673119**, and the probability of the serving side losing the point is **0.326881**. Conducting a non-parametric test to check if the probability of the serving side winning is significantly larger than the probability of the receiving side winning.

$$H_0 : p_{sw} = p_{rw} = 0.5 \quad H_a : p_{sw} \neq p_{rw}$$

$$n = 7284, \quad \mu = n \times p_{sw} = 3642, \quad \sigma = \sqrt{np_{sw}(1 - p_{sw})} = 42.6732$$

$$P(\text{4903 or more serving winning points}) = P(X \geq 4902.5) = P(z \geq \frac{4902.5 - 3642}{42.6732}) \approx 0$$

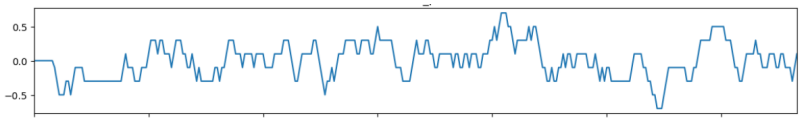
p-value is rather small, we can accept the hypothesis null that the probability of serving player winning is significantly larger.

Description of the State Space in a Time Series

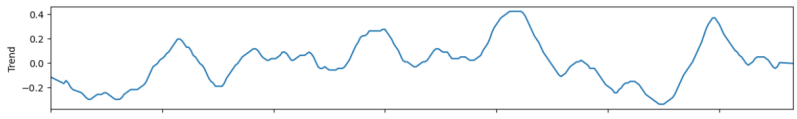
We formulate an equation to describe the dynamics of score changes. There are two main factors that determine whether a player can win this point: **serving or receiving** (as we've discussed) and **player's real-time performance status**. We use the additive model of state space to capture this relationship.

$$y_t = S_t + P_t$$

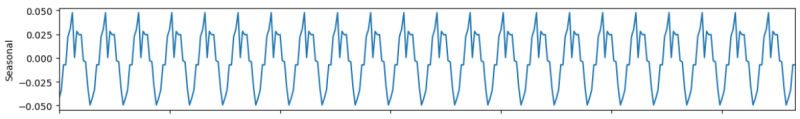
y_t denotes the player's current scoring flow, S_t denotes the impact of serving and receiving, and it is seasonal, P_t denotes the player's current performance status flow. The figures below visualize the equation. The x-axis represents the point number in a tennis match. The y-axis in figure y_t represents the net pointing rate in the last 10 points. The y-axis in figure P_t represents the player's performance value. The y-axis in figure S_t represents the impact of serving and receiving (positive values represent serving, while negative values represent receiving serves).



y_t



P_t



S_t

File" Wimbledon_featured_match" store the data and point change y_t and server S_t. Our target is to remove S_t from y_t to derive serial data that capture players' performance.

SARIMA Model

SARIMA is a time series forecasting model that combines elements of auto-regressive integrated moving average (ARIMA) with seasonality components. [8][9]

ARIMA model:

$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \Phi_1 \epsilon_{t-1} + \Phi_2 \epsilon_{t-2} + \cdots + \Phi_q \epsilon_{t-q}$$

The SARIMA model is an extension of the ARIMA model. It builds upon the ARIMA model by introducing three additional hyperparameters to define the autoregressive (AR), differencing (I), and moving average (MA) components of the seasonal aspect in the time series.

$$SARIMA \underbrace{(p,d,q)}_{non-seasonal} \underbrace{(P,D,Q)_s}_{seasonal}$$

p (AR order) denotes the number of autoregressive terms. **d (Integration order)** denotes the number of times the time series needs to be differenced to achieve stationarity. **q (MA order)** denotes the number of moving average terms. **P (Seasonal AR order)** denotes number of seasonal autoregressive terms. **D (Seasonal Integration order)** denotes the number of times the seasonal differencing needs to be performed. **Q (Seasonal MA order)** denotes the number of seasonal moving average terms. **s (Seasonal period)** denotes the number of time periods in each season.

First, we need to determine the key parameters in the model. As we've discussed earlier, players take turns to serve which is quite similar to the seasonal factor. Now we have to determine the s seasonal period of the serving and receiving circle. We assume that The number of rounds in each game is stable. The data set in "Wimbledon_featured_match" records a total number of **258** games and a total number of **1577** points from the quarterfinal. Average point rounds per game is **6.11**. Thus, a serving game and a receiving game form a set, and the average number of rounds in this set is **13**. Therefore, set s (Seasonal period) as **13**. Other parameters (p, d, q, P, D, Q) are more complicated to determine which we will discuss in the following subsection.

p	q	d	P	Q	D	s
TBD	TBD	TBD	TBD	TBD	TBD	13

Besides, we are required to input time series data y_t to derive fitted values after detrending and removing the seasonality. Here y_t represents point variation. As we are required to depict real-time performance fluctuation, we focus on net points scored over a certain period of time rather than the overall points. We have set up some variables to describe:

- npt_t: net point scored. npt_t = player1 points scored − player2 points scored
- $rnpt_{t,w}$: rolling net point scored, representing net point scored in a given period **w**

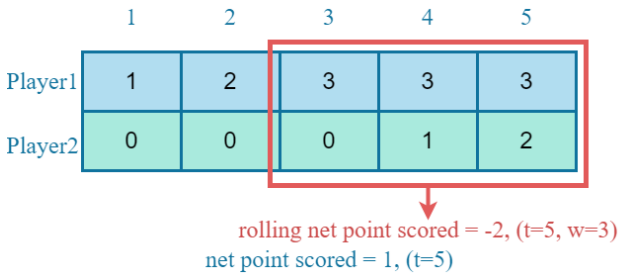


Figure above is a visualized illustration of npt_t and $rnpt_{t,w}$. $rnpt_{t,w}$ describes the real-time points flow using a rolling window to compute points difference in recent rounds. Therefore, $rnpt_{t,w}$ is time series data y_t we need.

By running the program, we compute all the values of $rnpt_{t,w}$ in every point of the match.

```
for match in wimbledon matches:
    rnpt = npt.rolling(window size).compute_differece
    save the rolling net point scored in a new column
```

Now that we've established the SARIMA model and derived the time series y_t, we can further our work to solve the model and depict players' real-time performance flow

5.3 Solving SARIMA Model

Time Series Data Stabilization

SARIMA requires that the time series is stationary. Stationarity helps ensure the robustness of the model's parameter estimation and predictions.

First, we conduct **ADF test**(Augmented Dickey-Fuller) to determine whether the time series is stationary. Here is the test results:

ADF statistics	-1.8577047921697456
p-value	0.35221203469395096

We can see that the p-value is relatively large, indicating that our data is not stationary. We need to conduct seasonal differencing to stabilize data. As we've discussed before seasonal period equals 13, set the parameters of `pandas.diff()` function as 13

```
import pandas as pd
read rolling net point and store it in a DataFrame "y"
y=y.diff(13)
```

Perform Augmented Dickey-Fuller (ADF) test on the differenced data. Here is the test results

ADF statistics	-5.379938568117287
p-value	3.73181071606e-06

We can see that the p-value is rather, indicating that we successfully removed the seasonality and stabilized the data.

Determine Value of Parameters in SARIMA model

$$SARIMA(p,d,q)(P,D,Q)_s$$

We've set s as 13 earlier. Now let's discuss other parameters. To determine the values of these parameters, we first need to plot the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the data.

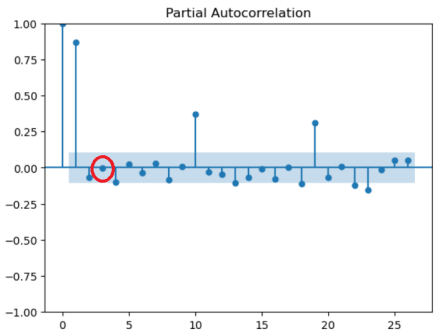


Figure 5: PACF

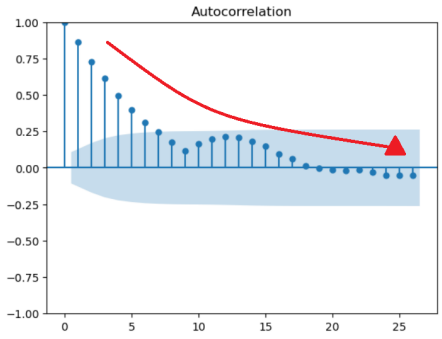


Figure 6: ACF

The following table illustrates how to determine the appropriate model and values for p and q by examining the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.

	AR(p)	MA(q)	ARMA(p,q)
ACF	Tail	Truncation after q order	Tail
PACF	Trunction after p order	Tail	Tail

In the PACF plot, we observe truncation after the 3rd order. In the ACF plot, we observe a trailing pattern.

To more accurately determine the values of p and q, we need to perform a grid search to find the values of p and q that minimize the **AIC**(Akaike Information Criterion). The objective of minimizing AIC is to maintain model fitting quality while minimizing the number of parameters in the model to prevent overfitting. We set **(p, q, P, Q)** to integer values ranging from 1 to 4, enumerate all possible combinations, and calculate the corresponding AIC values. The following table illustrates the results of our grid search to minimize AIC.

	(p, q, P, Q)	AIC
1	(3, 2, 0, 3)	-262.711646
2	(3, 2, 0, 2)	-258.608367
...	...	...
256	(1, 0, 0, 0)	144.780937

(3, 2, 0, 3) minimize the AIC. Besides, we performed one seasonal differencing, so D=1. In conclusion,

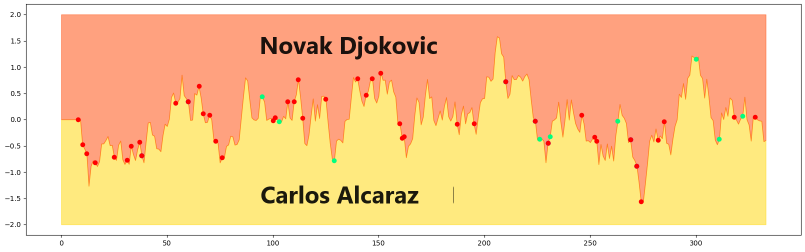
$$SARIMA(3,0,2)(0,1,3)_{13}$$

And then fit the model. The fitted value in this model depicts tennis players' performance

```
from statsmodels.tsa.statespace.sarimax import SARIMAX
model = SARIMAX(y, order=(3, 0, 2), seasonal_order=(0, 1, 3, 13)).fit()
Visualize fitted value of model
```

5.4 Visualizing Performance and Model Evaluation

We select data from the 2023 Wimbledon Man's Single Final for experimentation. The below figure visualizes players' performance status.



The yellow part represents Player 1, Carlos Alcaraz, Red part represents Player 2, Novak Djokovic. **x** axis is the point number of the match, which has a sequential pattern. **y** axis measures how well a player performs compared to his opponent. For example, **y** equals 0, indicating the two players are evenly matched. **y** larger than 0, indicating player1 performs better than player2.The larger the y (the area of the yellow region), the better the performance of Player 1. Vice Versa.

Does this chart accurately represent the players' states? We will compare it with the actual situation of the 2023 Wimbledon Men's Single Final to assess the accuracy of the model. In reality, Novak Djokovic dominated the early stages of the match. As reflected in the chart, during the initial stages, the y value is less than 0, and the area of the yellow region is smaller than the red region. Later, Alcaraz staged a comeback, overpowering Djokovic in the next two sets. In the middle part of the chart, the y value is greater than 0, and the yellow area is large. In the fourth set, Djokovic regained his momentum, hence the appearance of another peak in the red region on the chart. Through the above experimental validation, we found that our model has high accuracy and truly reflects the fluctuations in the game's states.

Apart from that, we also plot some green and red dots on the plot. **Green dots** represents player1's good shot that can boost morale, including ACE, breakpoint winning, extended rallies winning, etc. **Red dots** represents player1's bad shot that may negatively impact player's state, including a double fault, unforced fault, etc. We can see that most of the **red dots** are distributed along a **descending line**, showing the connection between bad shot and decline in performance. And most of the **red dots** are distributed on turning point upward, showing that a good shot can significantly boost a player's state. These results align with our intuitive understanding.

5.5 Residual Analysis

Next, we will perform a residual analysis to check whether our model fits well. Figures below are "Standardized residual for 'r'", "Histogram plus estimated density", "Normal Quantile-Quantile

Plot" and "Correlogram"

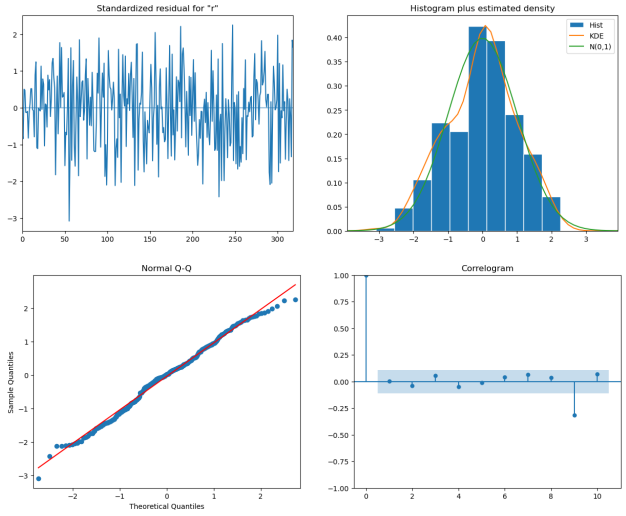


Figure 7: Plot Diagnosis

The QQ plot of residuals shows a line close to a straight line, indicating that the residuals are very close to a normal distribution. In the bottom-right corner, the residual autocorrelation plot also indicates the absence of significant autocorrelation in the residuals. This indicates that our model fits well.

6 Task 2: Demonstrate Existence of Momentum in Tennis Match

6.1 Problem Analyses

In this task, we are required to respond to coaches' doubts about whether or not momentum plays a role in tennis matches. In other words, when a player has scored consecutively, they are more likely to continue the momentum and score again. Conversely, the opposite is also true. Thus, we will discuss the correlation between the player's current status and the player's previous status.

In task 1, we've built a SARIMA model, which is a combination of AR(Autoregressive) and MA(Moving Average). The autoregressive model takes into account the correlation between the current state and previous states. In that case, we only need to explore the significance of the autoregressive coefficients in the model for Task 1.

6.2 Significance of Momentum

In Task1, we've built a SARIMA model:

$$y_t = \underbrace{\sigma + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \Phi_1 \epsilon_{t-1} + \Phi_2 \epsilon_{t-2} + \cdots + \Phi_q \epsilon_{t-q}}_{\text{non-seasonal}}$$

$$\underbrace{+\alpha_1 y_{t-s} + \alpha_2 y_{t-2s} + \cdots + \alpha_P y_{Ps} + \phi_1 \epsilon_{t-s} + \phi_2 \epsilon_{t-s} + \cdots + \epsilon_{t-Qs}}_{\text{seasonal}}$$

where $p = 3$, $q = 2$, $P = 1$, $Q = 3$, $s = 13$.

Based on the code for the previous model, print out the summary of the model. The following table reveals the significance level of the coefficients in the model.

	coefficient	standard error	p-value
AR. L1	0.6501	0.469	0.165
AR. L2	0.7895	0.350	0.024
AR. L3	-0.5583	0.274	0.041
MA. L1	0.2846	0.484	0.556
MA. L2	-0.5927	0.351	0.091
MA. S. L13	-1.6686	0.121	0.000
MA. S. L26	0.6194	0.137	0.000
MA. S. L39	0.0718	0.068	0.290
σ	0.0196	0.003	0.000

Table 2:

Clearly, we can see that $\beta_3, \alpha_1, \alpha_2, \alpha_3$ are significantly different from zero. This indicates that the player's performance is highly related to the previous 3 points as well as the previous 1 set. In conclusion, we are confident to claim the significance of momentum and rebut the coach's suspicion

7 Task 3: Predicting Changes in Momentum with Random Forest Model

7.1 Problem Analyses

Coaches want to know if there are metrics that can help determine when a game situation shifts from one player's favor to another's. To solve this problem, we need to build a predictive model based on historical game data, observe fluctuations in player momentum, and identify some of the most relevant shifting factors.

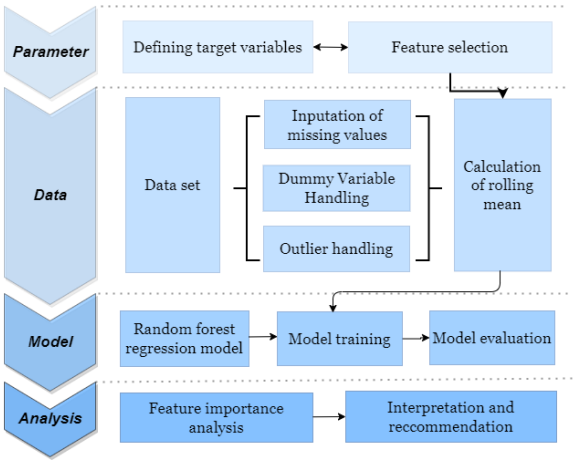


Figure 8: Flowchart for Task 3

7.2 Data Reprocessing

Feature Selection and Acquisition

Based on our search for expertise in tennis as well as psychological research, we identified the following 13 areas of possible influencing features:

Table 3: Influencing factors and their meaning

Characteristic	Definitions
hb_1_player1	Whether the last game was a break of serve
hb_2_player1	Whether the last game was a hold of serve
hb_3_player1	Whether the last serve was broken
hb_4_player1	Whether it is the first game
rolmean_ptdif	The impact of previous points
rolmean_speed	Rolling average of serve speeds
bp_effect	Points scored on break points=sum(break pt won) - sum(break pt missed)
ace_effect	The impact of ace ball
df_effect	The impact of double fault
unferr_effect	The impact of unforced erro
rundistance_effect	Rolling average of distance run
winner_effect	The impact of direct points scored by a player in a match
net_effect	Impact of net point

"hb" features(Hold& Break) are one-hot code measuring the impact of previous game. As the impact of serving games and receiving games on the match is different, we should identify whether the previous game was a receiving game or a serving game and discuss separately. There are total of foul cases:

Case	one-hot code
First game of the set	(0,0,0,1)
Win the previous serving game	(0,0,1,0)
Lose the previous serving game	(0,1,0,0)
Win the previous receiving game	(1,0,0,0)
Lose the previous receiving game	(0,0,0,0)

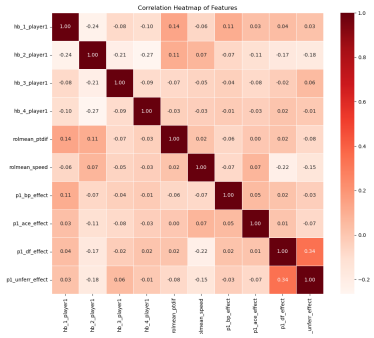
Table 4: Caption

Features "rolmean_ptdif", "rolmean_speed", "rundistance_effect" are a rolling mean of net point scored, serving speed, run distance in previous 7 rounds. Feature "ace_effect", "df_effect", "unferr_effect", "winner_effect", "net_effect" count the number of ace, double fault, unforced fault, winner point and net point in previous 7 rounds.

In particular, we focus on player's real-time index rather than data of the whole match, for the reason that it can't give a good representation of the momentum swings in the game over a short period of time. By trialing, we set rolling(window=7), which will lead to a better accuracy of the model.

Then, we performed missing value padding on the processed features. In this case, "rolmean_speed was padded using the mean value, while the other features were padded with zero.

Exploring Data Analysis



This heatmap shows the correlation coefficients between each pair of feature values. Clearly, we can see that most of the correlation coefficients are near 0, indicating that there is not a strong correlation between different features. That our model does not exhibit multicollinearity, which significantly improves the stability and accuracy of the model.

Feature Engineering

We randomly split the dataset into training and validation sets, and normalize the data(this is not necessary in our random forest regression model)

7.3 Random Forest Regression Model

7.3.1 Establishment of the Model

The random forest regression model is an integrated learning method that combines the ideas of bagging (bootstrap aggregation method) and random feature selection to perform regression tasks by **integrating multiple decision trees**, which can be relatively good predictors of fluctuations in the match situation. **It can handle large-scale data and reduce the risk of overfitting.** Random Forest has **greater robustness**. It is relatively less susceptible to outliers and noise because it is based on the combined opinions of multiple decision trees rather than relying on a single model. We will demonstrate this later in the model analysis and validation section.[10]

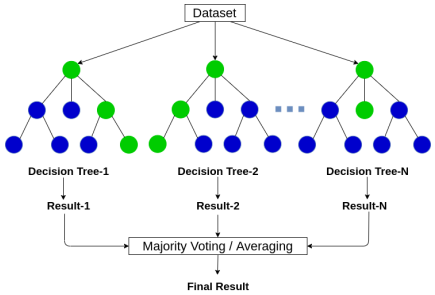


Figure 9: Integrated Learning for Random Forests[11]

We created a random forest regression model, set parameters such as the number of decision trees, maximum depth, etc., trained the random forest regression model using the training set X_train and the corresponding target variable y_train, and made predictions on the test set.

```
rf_regressor = RandomForestRegressor(n_estimators max_depth)
rf_regressor.fit(X_train, y_train)
y_pred = rf_regressor.predict(X_test)
```

Our Random Forest Regressor predicts the probability of winning each point based on input features. In next section **Task4**, we will visualize our predicition results and evaluate our model by experiment and measuring some key values, including R^2, recall, precision,F1, etc.

7.3.2 Sensitivity Analysis & Feature Importance Detection

Now, we perform sensitivity analysis and feature importance detection to determine which detail affects the match most. Bases on that, we give coaches advice to better prepare the match. We extracted the importance of features from the trained Random Forest regression model by traversing the feature names and printing the feature importance.

Table 5: The importance value of the feature

Feature	Importance Value
hb_1_player1	0.0009572459622893
hb_2_player1	0.2979262452917718
hb_3_player1	0.0796260709227249
hb_4_player1	0.0010893971614877
rolmean_ptdif	0.0420147212513150
rolmean_speed	0.0288525406155366
bp_effect	0.0542358450898098
ace_effect	0.0125027329610564
df_effect	0.0002812711806832
unferr_effect	0.1753468791095946
rundistance_effect	0.0310233653541932
winner_effect	0.2465149276935555
net_effect	0.0296287574059814

Subsequently, we select those features whose importance is greater than 0.01 (which is the threshold we set) and treat them as higher-importance features, reinitialize and train a new random forest regression model using these selected features, and make predictions on the test set. We find that the predicted results are similar to the previous one, suggesting that those characteristics we ignored are indeed unimportant. The more specific values and visualization will be explicated in the Model Evaluation part.

We represent the importance of the remaining characteristics in a pie chart:

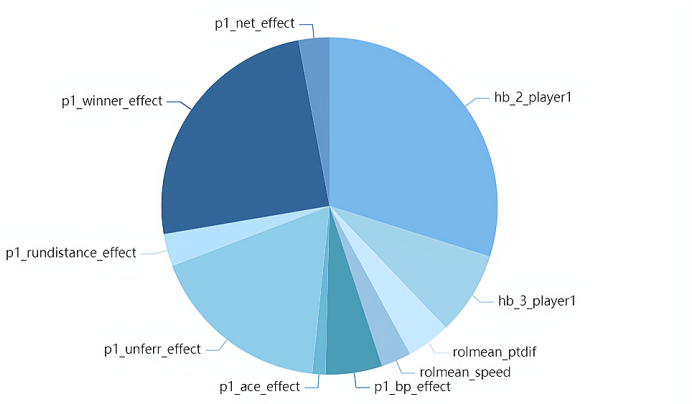


Figure 10: he weight of potential characteristics

From the graph, we can see that, hb_2_player1, p1_unferr_effect, and p1_winner_effect, these three characteristics have a significantly greater impact than the others.

7.4 Advice for Players Based on the Fluctuating Momentum

The use of hard power and tactical strategies are equally important in a match. To optimize player performance, we built models based on a large amount of data and identified key factors that affect momentum, such as a player's winning points, unforced errors, and whether or not a player breaks or holds serve in the final game of a match.

Utilizing these findings, we provide players with tactical advice when playing against different players. Players and coaching teams should consider various factors and develop a game plan that addresses the characteristics of the opponent (e.g., the success rate on breakpoints). For example, when facing opponents with a high capacity for winning points, players can adopt a more robust defensive strategy that focuses on consistency and reliability of returns, while when facing opponents with a high rate of unforced errors, they can try to induce their opponents to make mistakes by using versatile tactics.

To enhance their momentum in the game, players should focus on improving their ability to score winning points, which requires strengthening technical exercises in training, such as forehand and backhand strokes and interceptions, as well as improving the intensity and accuracy of their strokes. In addition, players should develop the ability to capture the moment in a match, initiate attacks, and win key points by hitting the ball accurately. It is also vital to maintain a stable mindset and minimize unforced errors during a match. Players should have the tactical flexibility to adjust their attacking and defending strategies, serving positions, and tactical combinations according to the match situation.

In conclusion, by combining data analysis and individual skill enhancement, players can better cope with different opponents, formulate effective tactical strategies, and maximize their strengths in matches to achieve better results.

8 Task 4: Evaluation of Random Forest Model

8.1 Test the Model with 2023 Wimbledon Final

We test our model in the match between Carlos Alcaraz and Novak Djokovic in the 2023 Wimbledon Gentlemen's final. We measured **MSE = 0.075** and **R-squared = 0.699**. The value shows that our regression model performs well.

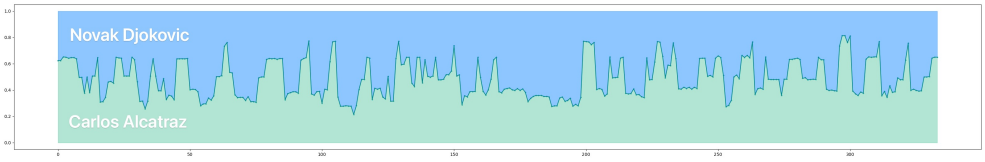


Figure 11: Fluatuating match situation in the final

We plot the predicted probabilities of the model using a scatterplot, connecting the lines and filling in the colors. The horizontal coordinate indicates the number of games played and the vertical coordinate indicates the probability of winning. Thus, the bottom green part of the image shows the momentum of player1 (Carlos Alcaraz) and the top blue part shows the momentum of player2 (Novak Djokovic).

The image visualizes the real-time status of the two players in the final, showing how the up-and-coming youngster broke Novak Djokovic's winning streak.

8.2 Model Accuracy Check

Confusion Matrix

Next, we use a confusion matrix to compare the relationship between the model's predictions and the actual labels and visualize the performance of the model. We use the model predicted probabilities stored by y_pred_selected and convert them into binary labels by thresholding 0.5, i.e., labeling probabilities greater than 0.5 as 1 and probabilities less than or equal to 0.5 as 0. So:

y_final: the actual target variable, representing the true category of the sample.

y_pred: the prediction result of the model, which is the target variable predicted based on the input features

The confusion matrix is a 2x2 matrix whose elements include:

- Top left (FN): False Negative
- Upper right corner (FP): False Positive
- Bottom left (FN): False Negative
- Bottom right (TP): True Positive

The ROC curve

We then evaluate the performance of our binary classification model using the ROC curve and the AUC value, with the false positive rate (FPR) as the horizontal axis and the true positive rate (TPR) as the vertical axis. The closer the curve is to the upper left corner, the closer the AUC value is to 1, indicating better model performance.

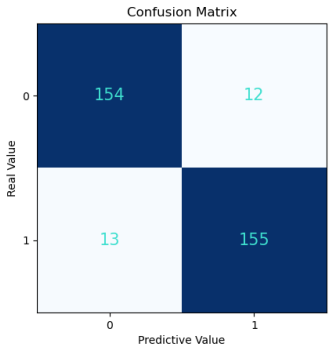


Figure 12: Confusion Matrix

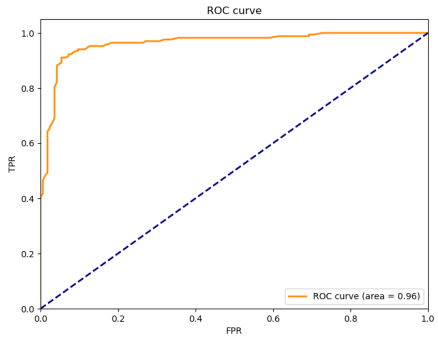


Figure 13: ROC curve

We show the comparison between the Real Value and the Predictive Value of the model through a scatter plot. The actual target variable is represented by the blue point and the model's predictive result is represented by the red point. We can see that the blue and red points in the scatterplot will tend to coincide, indicating that the model's predictive value is almost the same as the actual value with high accuracy.

Figure 14: Scatterplot of real and predicted values

Other Model Evaluation Metrics:

Table 6: Four model evaluation metrics

Evaluation metrics	Value
Accuracy	0.9281437125748503
Precision	0.9235294117647059
Recall	0.9345238095238095
F1 Score	0.9289940828402367

We also tested the generalization ability and robustness of the model using **cross-validation**:

R-squared obtained by 5-fold cross-validation: **0.709710, 0.706088, 0.727912, 0.752583, 0.718105**

Through the above evaluation and test, we can prove that the classification ability and generalization ability of our model are good and the accuracy of the model is rather high.

8.3 Model Generalisation Check

Now that we have proven that the model is accurate and effective, we then want to know how well it performs when detecting different types of events (such as women's matches, championships, etc.).

Reasons for our detection:

The model training process mainly uses data from men's games. If the data distribution in other contexts differs significantly from the training data, the model may not generalize accurately.

For example:

- The men's and women's competition formats differ in strength and speed, running range, emotions, and psychological states.
- Different tournaments may have different rules, schedules, and environments.
- The court surface (e.g. grass, soil, hard surface) may affect a player's performance.
- Different types of sports can have very different rules, techniques, and tactics, causing the model to fail.

So we looked for the data of the Australia Open Women's Single on the official website and adjusted the data format to be the same as the data given in the title, mainly by adjusting the naming of the first row of each column to be the same. The data cleaning and feature processing part is the same as the previous method. We also evaluated and tested our model with the first match of the semi-final between player Carlos Alcaraz and player Daniil Medvedev(the data was already given in the topic).

Then we the trained model to predict and evaluate the new matches' data and get the following values:

Table 7: Metrics for evaluating model performance

Competition type	MSE	R-squared
The first game of the semi-final	0.06509209807881693	0.7340670117112846
Australia Open Women's Single	0.0711407685247699	0.7131158960856507

The results we obtained are satisfactory, indicating that the generalization of the model is very good.

8.4 Adjustments for Future Models

If the model performs poorly at times, some potential factors might need to be included in future models.

- **Match Stage Factors:** Consider including the stage of the game, e.g. whether the game is in the preliminary or final stage, which may affect the performance and motivation of the players.
- **Sample size factors:** It may be that the sample size is not sufficient, different competitions have different rules, schedules, and environments, and these factors may have an impact on the model's predictions.
- **Opponent Factors:** Consider factors that encompass the level of opponents, previous performances of opponents, historical meetings with opponents, etc. Players may perform differently against opponents of different levels.

We can consider these factors in future models as needed to improve their performance. To improve the model's ability to generalize across different scenarios, we can use the following approaches:

- **Richer Data**: Ensure that the training data covers a wide range of different scenarios to better capture variations. The future models can be trained with data from men and women, singles and doubles, of different types of sporting events, to improve the robustness and generalization of the model.
- **Feature Engineering**: Perform feature engineering for different match types and scenarios to enable the model to better understand and utilize key features.
- **Domain Knowledge Incorporation**: Incorporate domain expertise to understand the specificities of different match types and scenarios to better tune the model.
- **Model Tuning**: Improve performance by adjusting the model structure, and hyperparameters, or selecting models that are better suited for multi-scenario prediction. When trying to adapt to different scenarios, it is key to closely monitor the model's performance and make timely adjustments.

Additionally, there are several other objective factors on the playing field, such as players' physical condition, injuries, playing field, and weather conditions, but these are difficult to quantify as model impact factors.

9 Strength and Weakness

9.1 Strength

1. **Scientific Indicator Selection**: We scientifically selected evaluation indicators by analyzing the correlation between 13 potential factors affecting game momentum. This involved utilizing a feature correlation heat map to enhance model prediction accuracy.
2. **Effective Model Performance**: Our SARIMA model, an extended ARIMA model, addresses time series data with significant seasonal fluctuations, ensuring accurate assessment of a player's true performance state. Additionally, a random forest model predicts momentum fluctuations with high sensitivity and generalizability, achieving almost all objectives.

3. **Robust Overfitting Resistance**: The random forest model, comprising independently trained decision trees with randomly selected samples and features, exhibits strong resistance to overfitting. This reduces dependence on specific data portions or features, enhancing the model's accuracy in prediction.

9.2 Weakness

1. **Interpretability Challenge:** The Random Forest model's intricate internal structure and algorithms pose challenges to understanding its prediction rationale due to complexity.

2. **Limited Model Validation:** We did not validate our model across various sports due to difficulties in obtaining detailed supporting data. Additionally, variations in rules and tactics across different sports may not align with the metrics established for our initial model.

9.3 Further Discussion

- **Model1. SARIMA**:

  To boost SARIMA model accuracy, integrate external factors (holidays, promotions) and regression variables. Explore model fusion with multiple SARIMA predictions. Implement autoARIMA or hyperopt for optimal configuration, and enhance accuracy by incorporating more historical data to capture underlying time series patterns and trends.

- **Model2. Random Forest Regression**: As we've discussed in **Task4** section, we can improve our model by increasing our dataset, feature engineering, selecting more related feature, etc.

References

[1] T. Coulon, H. Barki, and G. Paré, "Conceptualizing project team momentum: A review of the sports literature," *International Journal of Managing Projects in Business*, vol. 14, no. 2, pp. 270–299, 2021.

[2] K. L. Burke, T. C. Edwards, D. A. Weigand, and R. S. Weinberg, "Momentum in sport: A real or illusionary phenomenon for spectators.," *International Journal of Sport Psychology*, 1997.

[3] S. Avugos, J. Köppen, U. Czienskowski, M. Raab, and M. Bar-Eli, "The "hot hand" reconsidered: A meta-analytic approach," *Psychology of Sport and Exercise*, vol. 14, no. 1, pp. 21–27, 2013.

[4] M. I. Jones and C. Harwood, "Psychological momentum within competitive soccer: Players' perspectives," *Journal of Applied Sport Psychology*, vol. 20, no. 1, pp. 57–72, 2008.

[5] R. J. Vallerand, P. G. Colavecchio, and L. G. Pelletier, "Psychological momentum and performance inferences: A preliminary test of the antecedents-consequences psychological momentum model," *Journal of Sport and Exercise Psychology*, vol. 10, no. 1, pp. 92–108, 1988.

[6] J. Taylor and A. Demick, "A multidimensional model of momentum in sports," *Journal of Applied Sport Psychology*, vol. 6, no. 1, pp. 51–70, 1994.

[7] B. Moss and P. O'Donoghue, "Momentum in us open men's singles tennis," *International Journal of Performance Analysis in Sport*, vol. 15, no. 3, pp. 884–896, 2015.

[8] A. K. Dubey, A. Kumar, V. García-Díaz, A. K. Sharma, and K. Kanhaiya, "Study and analysis of sarima and lstm in forecasting time series data," *Sustainable Energy Technologies and Assessments*, vol. 47, p. 101 474, 2021.

[9] G. Kirchgässner, J. Wolters, and U. Hassler, *Introduction to modern time series analysis*. Springer Science & Business Media, 2012.

[10] X. Li *et al.*, "Using" random forest" for classification and regression.," *Chinese Journal of Applied Entomology*, vol. 50, no. 4, pp. 1190–1197, 2013.

[11] A. Sharma, *Decision tree vs random forest: Which algorithm to choose*, `https://www.jiqizhixin.com/articles/2020-06-11-6`, 2020.