Contents

[**Executive Summary** 2](#_Toc174570496)

[**Acknowledgment** 3](#_Toc174570497)

[**Introduction:** 4](#_Toc174570498)

[Background 4](#_Toc174570499)

[Motivation 4](#_Toc174570500)

[Research Objective and scope of project: 5](#_Toc174570501)

[**Literature Review** 6](#_Toc174570502)

[Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments 6](#_Toc174570503)

[Visualizing Professional Tennis Upsets: ATP 2012-2014 Men's Singles Matches 6](#_Toc174570504)

[Tracking the Evolution of a Tennis Match Using Hidden Markov Models 7](#_Toc174570505)

[Data-driven analysis of point-by-point performance for male tennis player in Grand Slams (https://revistas.rcaap.pt/motricidade/article/view/16370) 8](#_Toc174570506)

[Capturing Momentum in Tennis (https://theanalyst.com/eu/2022/03/capturing-momentum-in-tennis/https://theanalyst.com/eu/2022/03/capturing-momentum-in-tennis/) 9](#_Toc174570507)

[Quantification of momentum in tennis matches and its impact: a study based on AHP-EWM method and data analysis 10](#_Toc174570508)

[Applications of Machine Learning to Optimize Tennis Performance: A Systematic Review 10](#_Toc174570509)

[References: 12](#_Toc174570510)

# **Executive Summary**

The primary objective of this dissertation is to broaden the understanding of the variables influencing professional tennis match results by exploring into the dataset of ATP tennis match statistics. This report looks at key performance measures like break point saving ratio, ace rate, and first point win % over several ATP seasons using the extensive Exploratory Data Analysis (EDA)technique. This Study involves having around ~ 4000 data points by extracting ATP matches from year 2019 to 2023 across tourneys of all levels.

Advanced statistical methods such as ANOVA and Tukey HSD were employed to examine surface-level data, resulting in substantial findings into the variations in player performance on various court surfaces. The study used Markov Chain modelling to extract hidden states from the performance measurements of the winner and loser in order to gain a deeper understanding of match dynamics and to illuminate the momentum shifts that occur during a match.

Leveraging historical match data, we attempted to implement a Classification machine learning model that predicted the result of the important third set, building on these fundamental studies. Building our understanding over several key match performance metrics and incorporating it by using relevant feature engineering techniques. The capacity to forecast the future may have significant effects on betting markets, in-game decisions and game strategy.

# **Acknowledgment**

# **Introduction:**

# Background

Ever since the modern days of sports have been integrated with the application of Data Analytics, every aspect of the game from players performance to team strategy, fan engagement to business model has been transformed into multiple folds. (Srivastava, 2023) .

According to a report by Fortune Business Insights, global sports analytics market has forecasted to reach form $3.78 billion in 2023 to $22.13 billion by 2030, showing a substantial growth of 28.7% CARG. One prime example for this is the sports of F1, which at current time is so data driven backed up by AWS, providing top notch viewing experience for the viewers by showing real time stats during race like tire deterioration, traction ratio , all the sector times with coloured mini sectors, too see where exactly a driver has improved his lap time and DRS status .

First official recorded match statistics for a Tennis match could be found back in 1991. Since then the hidden complexities associated with the game of tennis has always be an challenge in the statistical analysis of the game. In late 2013, Novak Djokovic followed by Federer in 2017 , data analytics was introduced . Tennis unlike other sports is a very individualistic game, hence the trend of advanced analytics was slow to gain traction. Drone technology and 3D printing is used in new tennis ball machine to generate a complex set of shots using a Smartphone to control speed and spins. Building new tennis metrics – Big technology players IBM, Infosys, and SAP are continually upgrading new tennis metrics. Drone technology and 3D printing are employed in a new tennis ball machine to manufacture a complex series of strokes, with a Smartphone controlling speed and spin. Big technology companies like IBM, Infosys, and SAP are constantly updating new tennis metrics. (Drive-analytics, n.d.)

## Motivation

As quoted by Serena Williams, professional tennis player *“Data science has revolutionized how we approach training and preparation. It provides us with valuable insights that help us optimize performance and reach our full potential”.* “MoneyBall”, a 2011 film based on the story of a baseball team, Oakland Athletics who changed the landscape of competitive baseball by leveraging Sabertmetrics. Not only they won 20 consecutive games but also did that but spending lowest cost per win than other games, as it started with mere 10000 data points which has now rose to a staggering 10 billion data points.( <https://bbiasblog.com/2022/12/10/moneyball-how-sabermetrics-changed-baseball-forever/>)

Tennis being a individualistic game, the competitiveness of the game keeps of increasing year after year. The relevance of data analytics in Tennis could be a deterministic parameter between winning or losing a crucial series. With young talents like Carlos Alcaraz emerging, EDA and advanced statistical models could be leveraged to recognise such early trends. Majority of top 20 ATP professional athletes have started to leverage the use of analytics. The adaptation of this technology is slow but incremental unlike in other sports mainly due to the nature of the game. Currently the main technique for analysis is Labelling the match videos to evaluate various player performance metric. Point progression in each set, momentum shifts, Break points analysis.

## Research Objective and scope of project:

Given the extensive scope of analysing ATP tennis match data, the main objective of this dissertation is to study the dataset and define particular research questions that can provide significant insights into the factors that influence match outcomes in professional tennis.

The primary purpose is to undertake a thorough Exploratory Data Analysis (EDA) on the most recent complete ATP season, with the goal of understanding significant statistics and visualising them. This preliminary research will help to find interesting features that can be examined further, possibly through the analysis of historical data from prior seasons.

This dissertation will investigate the following research directions:

1. Examining "Upset" Matches: Using the match data, the study will try to pinpoint the salient features or traits of "upset" matches—those in which a player with a considerably lower ranking than the opponent wins—in order to arrive at conclusion. This will entail establishing the parameters for what qualifies as a "upset" and examining the variations in match statistics between upset and non-upset contests.
2. Comparing Match Statistics Across Eras: This research will examine the shifts in match statistics in professional tennis over various eras by utilising match data from several ATP seasons. This comparative study can shed light on how the sport's dynamics, tactics, and general level of competitiveness are evolving. Deeper insights into the observable traits will be found by using advanced statistical techniques like ANOVA and Tukey HSD.
3. Third Set Match Outcome Prediction: As an attempt to provide further insights, dissertation will explore developing a machine learning model to predict the winner of the third set in a match. Historically, the momentum and final result of the match have been greatly influenced by the third set. In order to calculate the likelihood that a player would win the crucial third set, the model will make use of players' meta data and performance measures from the earlier sets.

With possible applications in player development, coaching tactics, and tournament preparation, the investigation of these study aims will deepen our understanding of the elements that impact professional tennis success.

# 

# **Literature Review**

## Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments

In the field of tennis match prediction, a groundbreaking study emerged that combined empirical data with expert judgments, achieving a remarkable 85.1% accuracy rate. This research utilized an extensive dataset from the ATP World Tour, spanning from 1968 to 2015, encompassing 165,974 data points across 44 features. The study employed an Analytic Network Process (ANP) model to predict match outcomes, focusing specifically on contests between Djokovic and Federer. To assess the significance of player performance metrics, the researchers conducted a Wilcoxon rank sum test on 20 key features. Linear regression was then applied to determine the relative importance of each feature through coefficient analysis.

In preparing data for the ANP model, the researchers innovatively grouped similar performance factors into clusters, effectively reducing the number of factors from 20 to 12. This clustering approach ensured that the model's inputs were both comprehensive and manageable. A key aspect of this study was its integration of expert opinions and pre-match judgments. These judgments were carefully constrained to be based solely on the features incorporated into the ANP model. Experts made paired comparisons drawing on their knowledge and experience, while also considering correlation coefficients from historical data. The researchers also factored in the current form of players and additional judgment opinions to create a prioritized list of influencing factors.

The ANP model's super matrix was then used to synthesize all these priorities, producing a final prediction. This methodology represents a significant advancement in tennis match prediction, combining the strengths of data-driven analysis with the nuanced insights of expert judgment. This approach not only demonstrated high predictive accuracy but also provided a framework for understanding the complex interplay of factors that influence tennis match outcomes. The study's success in integrating quantitative data with qualitative expert knowledge opens new avenues for sports analytics and predictive modelling in tennis.

## Visualizing Professional Tennis Upsets: ATP 2012-2014 Men's Singles Matches

This blogs by Tyler Knutson (NYC Data science academy) walks though the analysis of performance metrics and Odds of upset matches; where an underdog player defeats a most likely a favourite / higher seed player. So, identify the winner of the match before the match, the concepts of “Odds” is leveraged, Odds greater than 2.0 denotes an underdog while odds less than 2.0 represents the favourite. A density distribution of break points in a upset match is studied to analyse the impact on break point metric. The dataset used in this is very rich in terms of the detail and the granularity, as it has metrics like no of backhands, forehands, volleys, how deep was service return etc. This mainly highlighted 2 main findings:

Firstly, studies have observed a tendency among bookmakers to allocate their resources disproportionately when setting odds. Bookmakers focus more intently on matches with less predictable outcomes, dedicating less time to precisely calibrating odds for matches with clear favourites. For instance, in matches with heavy favourites, the difference between setting underdog odds at 20:1 versus 18:1 was often neglected.

Secondly, analyses of betting patterns across multiple sportsbooks have revealed that odds are frequently adjusted not just based on perceived probabilities, but also in response to the distribution of wagers. In a notable example, when two sportsbooks initially offered identical odds (1.3 for the favourite, 3.5 for the underdog), the book receiving heavier action on the favourite adjusted their odds to 1.2 and 4.5 respectively. This adjustment serves to encourage more balanced betting, allowing sportsbooks to manage their risk exposure more effectively.

The logit function and its applications in sports modelling

This (Hyatt-Twynam, 2013) blog post provides in-depth exploration of logit function and its applications in sports application. Logit function can be used as transformation while working with probabilities. Linear shifts in logit space are the fundamentals behind logistic regression. The author presents a compelling case for the utility of logit function in transforming probabilities to improve modelling accuracy and interpret uncertainties. The author argues that linear shifts in probability space can lead to unrealistic outcomes, especially in extreme probability scenarios. The logit transformation is presented as a solution, allowing for more sensible linear adjustments in logit space before converting back to probabilities.

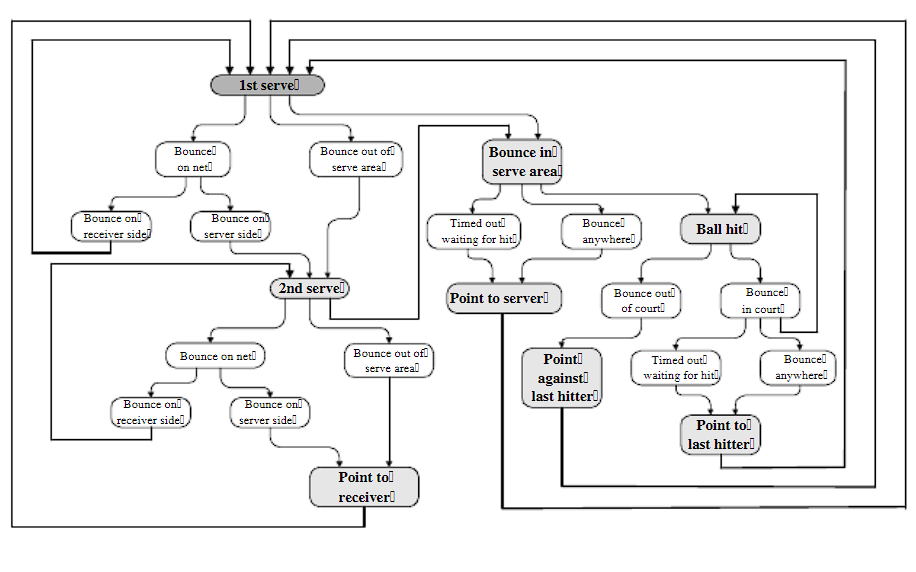
The blog post effectively demonstrates the versatility of the logit function in sports modelling, from basic probability transformations to more complex applications in Bayesian inference and market pricing. While the author provides personal insights and suggestions based on industry experience, the post does not reference specific academic literature or industry studies. This limits its academic rigor but enhances its practical relevance for those working in sports analytics. The author emphasizes the critical importance of quantifying uncertainties in probability estimates for both bookmakers and sports traders. Instead of relying on a single probability value, the blog suggests developing a probability density function (pdf) that describes the likelihood of each possible true probability value.

The blog introduces the logit-normal distribution as an ideal choice for representing this uncertainty. This distribution is created by applying a normal distribution in logit space. The author states: "Typically in science, for physical measurements, the logical choice would be a normal distribution, however if this is used for p\_A, you're claiming there's a chance p\_A could be outside the range 0 to 1. Once again, logit space comes to the rescue; using a normal distribution in logit space (a logit-normal distribution) for your prior solves this issue, and leads to a narrower distribution at the extremities."

## Tracking the Evolution of a Tennis Match Using Hidden Markov Models

The author IIias Kolonias, William Christmas and Josef Kittler has addressed the problem of extracting higher level semantic information from low level feature from multimedia content. Feeding a sequence of video as input data in hierarchical structure consisting of Hidden Markov Model. A graphical model is developed to represent the usual progression of the game, which helps to identify and examine elementary events within the tennis sequence. Using this fundamental sequence of events, reasoning for higher events is performed such as awarding the current point. Computer Vision models is used to detect the event and HMM is used to perform higher level reason.

This (Ilias Kolonias, August 2004) works aims in substituting the original scene evolution model with a series of smaller models, each of which aims to accurately depict a particular situation in the match's history. The most crucial thing we must make sure of during this process is that, upon combining every model in this set, we will have a model that is identical to the original.



**1 Graphical model for awarding a point in a tennis match**

Modelled events of lower conceived importance through an HMM, which would then trigger another HMM to infer on more important events within the game; that would also help us prevent spurious data from low-level feature extraction modules from propagating to higher levels of the inference engine

## Data-driven analysis of point-by-point performance for male tennis player in Grand Slams

This (Yixiong Cui, 2019) article illustrates the work on analysing the performance metrics of tennis by incorporating 29675 data points through Classification Tree analysis. The author leveraged the point level data of 145 Grand slams main draw men’s Singles matches played by top ranked players from year 2011 to 2016. A two-step cluster analysis was performed with Euclidean distance as measure and Schwart’s Bayesian criterion. The classifying metrics were (i) Serve speed (ii) Rally length.

This article projects the point level analysis of players using classification tree model. Humidity and Ball types were 2 unique features which could be found being used by the author. Another key finding was due to the slow court surface of Roland Garros, it was easier to break to the server’s game and players tend to player more aggressively. The influence of Rally points was noticeable in certain ways: Short and longer rally helped player to win more matches than medium rally . The serving efficiency of a player is not only affected by the surface type and the level of the tournament but also by the skill level of the opposition

In a thorough examination of tennis performance throughout Grand Slam tournaments, the author discovers that a classifying tree model portrays some critical insights concerning a player's effectiveness. The model shows that the analysed player had higher overall point-winning percentages in the Australian Open (AO) and US Open (US) than Roland Garros (RG) and Wimbledon (W).

Specifically, the tree model shows that serve speed has a considerable impact on point-winning chance. The player had a greater success rate with rapid serves throughout all events, with a clear advantage in the AO and US. Slower serves, on the other hand, resulted in decreased point-winning percentages, which were most noticeable in the AO and US. Interestingly, it demonstrates consistent return efficacy throughout all Grand Slams, with the player winning roughly one-third of points against first serves and 40% against second serves, regardless of event. This classification approach offers a comprehensive understanding of how surface qualities and serve dynamics influence match outcomes, which is useful for strategic planning and performance analysis in professional tennis.

## Capturing Momentum in Tennis

This (Manuel, March 4 2022) article sheds light upon identifying decisive momentum shifts moments in a tennis match and how such moments can be influential in loosing and winning the match. Leverage and Momentum are the 2 new Stats performance metrics which was introduced to identify key moments in WTA matches . Leverage measures the importance of a single point to the final outcome of a tennis match by quantifying how much a player’s probability of winning the match changes. Momentum aims to describe which player is in control at any point of the match

In a comprehensive study of tennis match dynamics, researchers have developed a sophisticated chain of predictive models to analyse and quantify the concept of momentum in professional tennis. The study, which utilizes data from 1.5 million points played on the WTA tour between 2012 and 2020, presents a novel approach to understanding the ebb and flow of tennis matches.

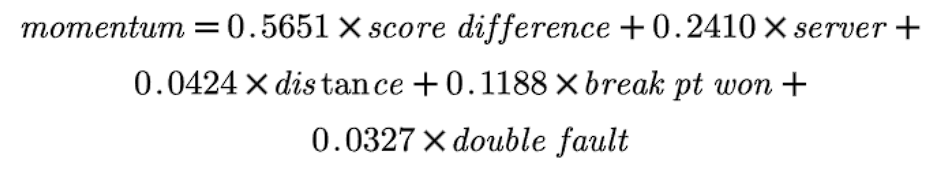
The authors introduce a multi-layered model that considers various factors including court type, current match state, in-match statistics, and pre-game odds. This model chain begins by predicting the probability of winning the next point, which then feeds into predictions for game, set, and ultimately match outcomes. This hierarchical approach allows for a granular analysis of how individual points impact the overall match probability.

Central to the study is the concept of "leverage," which quantifies the importance of each point in terms of its potential to change the match outcome. Building on this, the researchers define "momentum" as an exponentially weighted moving average of the leverage gained by a player. This definition takes into account both the recency and the importance of points, with more recent and higher-leverage points having greater influence on the momentum score.

The (Manuel, March 4 2022) study introduces the concept of a "momentum swing," defined as a shift in momentum from one player to another by a magnitude of 3% or more. This threshold provides a concrete metric for identifying significant turning points within a match.This research represents a significant contribution to the field of sports analytics, offering a data-driven approach to quantifying the often-intangible concept of momentum in tennis. The model's ability to consider a wide range of factors and provide point-by-point analysis of match dynamics offers potential applications for players, coaches, and analysts in developing strategies and understanding match progression.

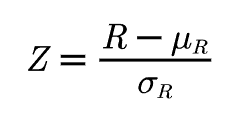
## Quantification of momentum in tennis matches and its impact: a study based on AHP-EWM method and data analysis

In this report (Wu, May 2024) , the author deeply analyses the concept of momentum formation in tennis match and its extent of influence on winning the match. Using Analytic hierarchy process (AHP) and the entropy weight method (EMP) methodology , Welch’s Test and Run test , non randomness of momentum metric is investigated. One interesting metric considered by the author during data pre-processing stage is p1\_d : the longer the distance traversed by an athlete, the greater their level of exhaustion consequently. Evaluating Criterion Matrix , final AHP and EWM weights are computed which are then fed into a concise formula to calculate momentum



2 Momentum formula

To establish the non random nature of momentum and likelihood of winning the match, Z test run is performed using below given formula:



Hypothesis test was undertaken on Wimbledon 2023 tournament – 1301 as reference for player Carlos Alcaraz . Z value obtained was -7.402, which indicate that w that the data are non-random data based on the variableMomentum\_p1, and the level is significant

## Applications of Machine Learning to Optimize Tennis Performance: A Systematic Review

This (Tatiana Sampaio, June, 2024) report sheds light over various approaches adopted to assess the performance stats of players . Notable reviewed mythologies include Psychological and affective states where a sensor based approach was employed to leverage inertial measurement units(IMUS) (Havlucu Hayati, 2022) worn by small group of elite coaches (n=2) and 4 professional players. BY using Long short term Memory Recurrent Neural Network, an accuracy of 85% was achieved. He further dwelled over the scope of AI for predicating players optimal performance “zones”.

This report also displays how various ML approaches yields different results for instance studies include Backpropagation and Neural Network and Convolutional Neural Networks being utilized to evaluate tactical performances and classifying groundstroke stances (Yu Zhang, 2023), (Yang, January 2023) .While Random Forest was utilized to perform Classification of moment patters and predicting point winners for a set. (Rosker Jernej, 2021) (Andrej Panjan, June 2010). (Yu Zhang, 2023) Designed a diagnostic model for assessing player performance using decision trees and a common classification technique while (Yang, January 2023) deployed convolutional neural network to analyse batting strength and angles based on video footage. CNNs excel at image recognition tasks, making them ideal for this application.

# Methodology

## Description of the ATP tennis match dataset

Here's a brief description of your dataset, categorizing the features into subcategories:

This dataset contains comprehensive information about tennis tournaments and matches, including player statistics and match outcomes. The data is organized into several categories:

1. Tournament Information:

|  |  |
| --- | --- |
| tourney\_id | A unique identifier for each tournament |
| tourney\_name | The name of the tournament, as various tourneys are played at different levels are listed in this : 470 different tourneys held from 2020-2023 |
| surface | The type of court surface , possible values are [Hard, Grass, clay] for year 2000-2003 era Carpet as a surface was also found |
| draw\_size | The number of players in the tournament |
| tourney\_level: | Indicates the level of the tournament (e.g., Grand Slam, Masters 1000, other level tourney, Challenger’s, Satellites/ITFs, finals, Davis cup) => Distinctive tourney levels values are [G,M,A,C,S,F,D] |
| tourney\_date | The start date of the tournament |

2. Match Information:

- match\_num: A match-specific identifier

- score: The final score of the match

- best\_of: Indicates whether the match is best of 3 or 5 sets

- round: The tournament round of the match

- minutes: The duration of the match

3. Player Metadata:

For both winner and loser:

- player\_id: Unique identifier for each player

- name: Player's full name

- hand: Player's dominant hand (right, left, or unknown)

- ht: Player's height in centimeters

- ioc: Player's country code

- age: Player's age at the time of the tournament

- rank: Player's ATP or WTA rank at the time of the tournament

- rank\_points: Player's ranking points

4. Match Entry Information:

For both winner and loser:

- seed: Player's seeding in the tournament

- entry: Type of entry into the tournament (e.g., wild card, qualifier)

5. Match Statistics:

For both winner and loser:

- ace: Number of aces served

- df: Number of double faults

- svpt: Total serve points

- 1stIn: Number of first serves made

- 1stWon: Number of first-serve points won

- 2ndWon: Number of second-serve points won

- SvGms: Number of service games

- bpSaved: Number of break points saved

- bpFaced: Number of break points faced

This dataset provides a rich source of information for analyzing tennis matches, player performance, and tournament characteristics. It allows for in-depth analysis of various aspects of the game, including serving performance, player rankings, and match outcomes across different tournaments and surfaces.

# References

Andrej Panjan, N. S. A. F., June 2010. Prediction of the successfulness of tennis players with machine learning methods. *Kinesiology.*

Drive-analytics, n.d. *The Gradual yet steady rise of Tennis Analytics.* [Online]   
Available at: https://www.drive-analytics.com/tennis.php

Havlucu Hayati, A. B. ,. E. T. ,. C. A. ,. O. O., 2022. Toward Detecting the Zone of Elite Tennis Players Through Wearable Technology. *Frontiers in Sports and Active Living,* Volume 4.

Hyatt-Twynam, S., 2013. *The logit function and its applications in sports modelling : Sports Trading Network.* [Online]   
Available at: https://www.sportstradingnetwork.com/article/logit-function-applications-sports-modelling/

Ilias Kolonias, W. J. C. J. K., August 2004. Tracking the Evolution of a Tennis Match Using Hidden Markov Models. *Lecture Notes in Computer Science.*

Manuel, J., March 4 2022. *Capturing Momentum in Tennis.* [Online]   
Available at: https://theanalyst.com/eu/2022/03/capturing-momentum-in-tennis

Rosker Jernej, M. R. Z., 2021. Skill Level in Tennis Serve Return Is Related to Adaptability in Visual Search Behavior. *Frontiers in Psychology.*

Srivastava, S., 2023. *appinventiv.* [Online]   
Available at: https://appinventiv.com/blog/data-analytics-in-sports-industry/

Tatiana Sampaio, J. P. O. A. M. P. N. E. M., June, 2024. Applications of Machine Learning to Optimize Tennis Performance: A Systematic Review. *Applied Sciences.*

Wu, W., May 2024. Quantification of momentum in tennis matches and its impact: a study based on AHP-EWM method and data analysis. *Highlights in Science Engineering and Technology.*

Yang, J. L. Z., January 2023. The Biomechanical Analysis on the Tennis Batting Angle Selection Under Deep Learning. *IEEE Access,* Volume 11.

Yixiong Cui, H. L.-Á. G., 2019. Data-driven analysis of point-by-point performance for male tennis player in Grand Slams. *revistas,* Volume 15.

Yu Zhang, 2023. The optimization of college tennis training and teaching under deep learning,. *Heliyon,* 10(4).