KTH ROYAL INSTITUTE OF TECHNOLOGY

Degree Project in Sports Technology

Second cycle, 30 credits

DIGITAL COACHING FOR TENNIS SERVE WITH MACHINE LEARNING

SERVE AND SHOT ANALYSIS AND GRADING FOR COACHING APPLICATIONS.

JORGE SÁNCHEZ DE LA VEGA ALCÁZAR

Master in Sports Technology

Date: August 26, 2024

Supervisors: Oscar Danielsson, Maksims Kornevs

Reviewer: Jonas Willén

Examiner: Sebastiaan Meijer

Swedish title: Digital coaching för tennisserve med maskinlärning

School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH

Trita number: TRITA-CBH-GRU-2024:336

Abstract

This project develops a novel approach to analysis and understanding of tennis shot and serve performance, with the help of machine learning for digital coaching purposes. Using the PlayReplay system, two models are developed to generate shot return probabilities and quantified serve quality. Shot return probability is used to provide insight into the decisiveness of a shot and its effectiveness. Built on that, serve quality is measured as the impact of such serves on subsequent shots, allowing for a more comprehensive understanding of serve quality. Key metrics for both shots and serves are computed on multiple samples obtained by rigorous data collection in other to maximize models' performance. Ultimately, this mentioned process allows the development of a final tool that offers specific feature analysis of a given serve, providing a suggestion on performance improvement of serve. The results show the potential of machine learning and data driven techniques to be implemented in real world scenarios and be used by players and coaches to improve player performance.

Sammanfattning

Detta projekt utvecklar ett nytt tillvägagångssätt för analys och förståelse av tennisslag och serveprestationer, med hjälp av maskininlärning för digitala coachningsändamål. Med hjälp av PlayReplay-systemet utvecklas två modeller för att generera skottretursannolikheter och kvantifierad servekvalitet. Sannolikhet för skottretur används för att ge insikt i ett skotts beslutsamhet och dess effektivitet. Baserad på det mäts servekvaliteten som effekten av sådana servar på efterföljande skott, vilket möjliggör en mer omfattande förståelse av servens kvalitet. Nyckelmått för både skott och servar beräknas på flera exempel som erhållits genom noggrann datainsamling i andra för att maximera modellernas prestanda. I slutändan tillåter den här nämnda processen utvecklingen av ett sista verktyg som erbjuder specifik funktionsanalys av en given serve, vilket ger ett förslag på prestandaförbättring av serven. Resultaten visar potentialen hos maskininlärning och datadrivna tekniker att implementeras i verkliga scenarier och användas av spelare och tränare för att förbättra spelarens prestation.

Key words

Digital coaching, tennis, serve, shot, sports performance, machine learning.

Nyckelord

Digital coaching, tennis, serve, skott, sportprestationer, maskininlärning.

Acknowledgements

I would first like to thank Jonas Willén for the support and guidance on my trajectory through the past two years of the Masters.

I want to thank Maksims Kornevs for the patience, understanding and commitment with all students and me, during the course of this thesis.

I want to thank Oscar Danielsson for his help, enthusiasm and believe in my ideas for this project, as well as PlayReplay for letting me work around the sport of tennis.

Thanks to the fellow players that participated in the data collection.

And lastly, special thanks to my parents, my brother and my friends, whose support has been key for me to feel comfortable in this new chapter.

Jorge Sánchez de la Vega Alcázar

Stockholm, August 2024

Content

1. Introduction

Tennis, as a sport, has continually evolved with advancements in technology. Among these innovations, the integration of camera-based systems on tennis courts has revolutionized the way the game is played, officiated, and analyzed, a well-known example is the addition of the Hawkeye system to the professional tour in 2006 [1]. These systems provide unprecedented levels of accuracy in ball tracking, player movement analysis, and statistical data generation. One such system, developed by PlayReplay [2], captures intricate details of a match, including ball trajectory, speed, spin, player positioning and several more.

Despite the variety of data these systems produce, there remains a critical need for a comprehensive analysis that translates this raw information into actionable insights for players and coaches on a commercial level. Specifically, the assessment and enhancement of a player's shots, present an opportunity for in-depth study and improvement. For instance, a 2020 study found tennis coaches spend their majority of work time in task that does not involve player coaching [3], Technologies as described could help with the requirement for man hours.

The primary aim of this project is to evaluate and analyze the performance of the tennis serve. Unlike other shots in tennis, the serve does not rely on the opponent's output but rather depends solely on one's own technique and skill. Given this unique nature of the serve, it becomes crucial to establish a method for determining the quality of a player's serve accurately.

The term “ace” serves as the ultimate classification for a well-executed serve. An “ace” is a serve that ends the point immediately, leaving the opponent unable to even reach the ball, thereby awarding the point to the server outright. While aces are straightforward to identify and serve as a significant metric for evaluating the performance of professional players, they present a more complex challenge when it comes to club-level tennis.

At the club level, aces are not as common, making them insufficient as the sole criterion for evaluating serve quality. This is because serves that do not result in aces can still be considered effective and valuable. Such serves can force the opponent to make a mistake or produce a weak return, thereby giving the server an advantage in the subsequent shot. Thus, it becomes essential to consider other factors and metrics to comprehensively assess the quality and effectiveness of serves beyond just the occurrence of aces.

1.1. Goals and Objectives

The goal of this degree project is twofold. Firstly, it seeks to harness the data output from the PlayReplay system to construct a robust model that evaluates a player's shot performance. This model analyzes key metrics such as ball speed, spin rate, placement accuracy, trajectory analysis and more, aiming to assign a quantifiable score to the shot. Such a scoring mechanism intends to provide a comprehensive view of the shot’s effectiveness, going beyond traditional statistics to offer a more nuanced evaluation. A similar system has been developed for professional level players by Tennis Insights, part of Tennis Data Innovations [4], to show tennis enthusiasts shot performance by their favorite players, however, there are no tools available for digital coaching at amateur level, as Tennis Insights focusses on generate statistics of professional level players for the sole purpose of visualization at consumer level.

For this last reason, the project attempts to utilize the generated shot scores to offer tailored recommendations for improvement to the players. By employing data-driven insights, this aspect of the research aims to provide actionable guidance, specific to individual players, to enhance their shot making capabilities.

Through the development of this scoring model and subsequent improvement suggestions, the project aspires to bridge the gap between raw data and practical utility, catering to the needs of both players and coaches. By providing quantifiable assessments and guidance derived from the data output of the PlayReplay system, this research proposes a contribute to the enhancement of player performance and the understanding of shot dynamics in the field of tennis.

1.2. Structure of the Thesis.

This thesis report will be structured the following way: (to be exactly defined)

- Background: The section discusses the technological advancements in tennis, particularly focusing on camera-based systems like PlayReplay that track ball trajectories and player movements. It examines how these technologies are transforming tennis by providing precise data for performance analysis, highlighting the shift from traditional coaching methods to more data-driven approaches. The background also touches on the limitations of current datasets available for tennis analysis and the need for more detailed data to fully understand the impact of serves on match outcomes.
- Methodology: refers to the use of the PlayReplay system to generate a unique dataset for analyzing tennis serves. It describes how the system captures 3D data using multiple cameras positioned around the court, and how this data is processed to track player positions and ball trajectories. The section also details the creation of two machine learning models, a logistic regression model for predicting shot return probability and a gradient boosting regressor model for scoring serve quality, used to evaluate serve effectiveness.
- Results: in this section, the results from the machine learning models are presented. The shot return probability model is evaluated for accuracy, revealing that serve speed, placement, and spin are critical factors influencing serve success. The serve quality model is analyzed for its performance, with discussions on the impact of dataset size on model accuracy. The results emphasize the effectiveness of these models in providing a more objective analysis of tennis serves compared to traditional methods.
- Discussion: interprets the results, focusing on the implications for tennis coaching and player development. It emphasizes the potential of data-driven insights to enhance training by offering personalized feedback based on specific serve characteristics. The section also explains the significance of this study in

sports technology, suggesting that the methodologies used could be adapted to other sports for similar performance evaluations.

- Conclusions: lastly, the report concludes by summarizing the key findings and their implications. It reiterates the value of using machine learning and detailed data analysis in understanding and improving tennis serve performance. The conclusion also highlights the potential for future research to expand on this work, exploring other aspects of tennis or applying similar techniques to different sports.

2. Background

This chapter include background of the project. Firstly, there is a need to assess the current technological landscape regarding the utilization of camera or sensor-based systems for tennis analysis, while also exploring their applicability to the sport. It also includes investigation into coaching methodologies emerged, focusing on whether coaches rely on tangible data or solely on empirical knowledge to instruct players the proper shot execution. Lastly, the availability of tennis datasets, particularly at point or shot level, is examined to ascertain their potential contribution to understanding the significance of serving in determining match outcomes.

Among the most notable advancements in professional sports technology is the Hawkeye system [1], initially developed for digital line calling. Evolving beyond its original porpoise, the system now facilitates real-time ball tracking and line calling. However, widespread adoption is hindered by stringent sensor requirements, limiting accessibility for many clubs. To address this challenge, ongoing developments such as PlayReplay [2], a camera-based system located at the net, are underway. Capable of ball tracking, line calling, shot recognition, and more, such innovations aim to democratize access to advanced analytical tools within tennis.

Historically, coaching techniques have relied heavily on non-written knowledge. However, with technological advancements, biomechanical studies, and performance analysis, a wealth of empirical data is now available. For instance, research conducted by the College of Sports and Exercise Science aimed to discern disparities in serve technique models between expert and novice coaches [5]. Similarly, studies from the

Universidad Complutense de Madrid have sought to develop models for providing feedback on serve quality based on questions answer by the performing player [6], and research by the Technical University of Munich aimed to understand serve placement dynamics and distribution [7].

While publicly available datasets for shot performance analysis are limited, notable exceptions such as The Match Charting Project offer valuable resources. Data from over 13,000 matches and nearly 8,000,000 shots recorded from professional players.

2.1. Tennis basics and terminology.

It is important to understand the basics of the sport of tennis and get familiar with the tennis terminology used in this report in order to follow the ideas presented in the following sections.

Tennis, as close as we know it today, is a sport racket invented in England at around the 16th Century [8]. Modern singles tennis is defined by the Cambridge dictionary as "a game played between two or four people on a special playing area that involves hitting a small ball across a central net using a racket" [9].

The mentioned special playing area is called the court, and it is divided in two sides, one for each player or pair of players; rectilinear lines define the court shape and size, an image of a tennis court and its dimensions can be found in Appendix A. Players can hit the ball before or after it has bounced once on their side of the court, if the ball first bounce occurs outside of the surface delimited by the lines the point is over, "point" being defined as the sequence of hits that bounce into the defined bounce area or that are hit before the ball hits the ground.

The action of hitting the ball towards the other side is named "shot" and it can by classified depending on its type:

- Serve: first shot of every point, required to be executed behind the farthest line of the bouncing area to the net, called the baseline. Players self-feed the ball in the air before hitting it and the bounce must land in a subsection of the bounce area representing approximately one fourth of a whole court side. Modern technique often requires the player to hit the ball over the head, and although not mandatory, this type of serve is analyzed in this project as it represents most of the serves performed at almost any level of the sport [10].

- Ground strokes: shots other than the serve which are hit after the ball has bounced. They can be divided into forehands and backhands: when the ball is hit on the side of the dominant or racket holding hand, it is considered a forehand, if hit on the opposite side it is considered a backhand.
- Volleys: shots other than the serve which are hit before the ball bounces, they as well can be classified as forehands or backhand.
- Returns: groundstrokes executed as the next shot after a serve, always hit after the ball has bounced.

In order to simplify the following reading of this report serves are referred to as "serves" themselves, and groundstrokes, including returns, and volleys are simplified as "shots", as only the differentiation of serves against not serve shots is needed.

2.2. Machine learning implementation.

Machine learning is a key component of a great quantity of technology currently developed and used. A variety of fields benefit from it and sports in one of them. The applications of machine learning in sports are immensely diverse, from result prediction [11] to injury prevention [12], athletes and teams from all types of disciplines take advantage of the current technology available to improve different aspects of their presence in the sport. Performance analysis and improvement being one of the biggest fields on which machine learning is relied on [13].

Formula 1, for example, known as one of the most technologically advanced sports, highly rely in artificial intelligence and machine learning to help teams make decisions in real live. An example of this technology is RaceWatch, racing performance analysis tool part of Catapult, company dedicated to sports performance analysis [14]. But not so technology focused sports also use machine learning.

IBM, company whose employee Arthur Samuel coined the term machine learning, defines it as "a branch of artificial intelligence (AI) and computer science that focuses on the using data and algorithms to enable AI to imitate the way that humans learn, gradually improving its accuracy" [15]. The benefit of machine learning is that, as it is performed by computers, the amount of data they are capable of processing is immensely bigger, making it key to solve problems humans are unable to solve due to complexity or size, as well as automatizing humans can regularly execute.

But there is a big array of different problems that can be solved by machine learning, and for that reason a variety of models exist. Each model has a distinct internal functioning structure in order to accommodate for the data used to solve a specific problem [16] [17].

This project relies on the use of machine learning to process high amounts of data to generate valuable and simplified outcomes that allow for a simpler understanding of the complexity of the tennis game. Two models are used for the current task:

- Logistic Regression model: logistic regression is one of the most used machine learning algorithms. Supervised algorithm widely used for binary classification tasks. Given a sample described as a set of features represented with values, a linear combination of said values is transformed into a probability of the sample to belong to either one of two defined categories, outputting a final value between 0 and 1 [18]. In the scope of this project, logistic regression is used to predict the probability of a tennis shot to be returned, furtherly explain in the following sections.
- Gradient Boost Regressor model: gradient boosting regression is a machine learning technique that constructs a robust predictive model by adding multiple weak models, typically decision trees. The process begins with an initial simple model and iteratively enhances it by training successive models to predict the residuals or errors of the current model. Each subsequent model addresses the errors of the one before, and this iterative improvement continues until the model's predictions achieve high accuracy. This methodology results in a powerful model by systematically correcting the inaccuracies of earlier iterations [19]. The result is a value representing the position of a sample in a given scale. A gradient boost regression model is implemented here to act as a score generator for a specific serve, depending on its characteristics, presented in next sections.

3. Methodology

3.1. PlayReplay System

To conduct a thorough analysis of the tennis serve, it is essential to have reliable data collection that the digital coach model can rely on. Unfortunately, for this specific project, there are no publicly available databases that meet the required criteria. It is challenging to find databases where shots are captured and measured using the same metrics as those used in this study, which will be discussed in more detail in the following sections of the methodology description.

There is one database worth mentioning, although it is not directly applicable to this project's development: The Match Charting Project [20]. This public and up-to-date database contains information from over 13,000 matches, with nearly 8,000,000 shots recorded. However, it does not offer the level of precision needed at the individual shot level for this analysis.

This lack of available datasets makes it imperative to create a specific, own, and unique dataset for the purpose of this project which is generated by the Play Replay system.

PlayReplay [2] is a Stockholm based company which develops a line calling system for club tennis courts that allows players not only to check and challenge calls (whether a shot was in or out), but also offers comprehensive tennis statistics for players from all levels.

Their system is based on the installation of four (or eight) cameras, two (or four) facing each side of the court, located on the top of the net posts. The cameras capture a wide image of the court, the ball, and the players, at a high sample rate. As two cameras are focusing on each side of the court, a 3D space can be triangulated from the images captured. Then, with the use of image analysis and computer vision algorithms both players position and ball trajectory are computed and tracked at different sample rates.

The system is also able to detect shots, shot types and shot outcome (in, net, out wide or out deep) and the position of the ball in key moments of the shot such as hitting instant, net crossing, and bounce.

Data outputted from the system is normally shown in an app, but in this case, as more precise information is needed compared to the one generated by the commercial application of the system, raw data regarding 3-dimensional ball and player position is also used, thanks to a modification of the standard firmware. The PlayReplay court layout and coordinate system is shown in Appendix B.

1. PlayReplay system: camera placement

3.2. Data collection

Data collections consist in playing sessions with full points, played in two venues: Salk Tennisklubb, located in Bromma, Stockholm; and Good to Great Tennis Academy, in Danderyd, Stockholm.

Four different males, right-handed players participated in the sessions throughout a 67-day period. Every player had several years of playing experience with an intermediate to advanced level.

Playing sessions are recorded by the PlayReplay system and videotaped separately, as a validation method and a helping tool in case of needed cleanup of the data from the PlayReplay system. The sensors from the system are capable of capturing ball positions throughout the entirety of gameplay, alongside tags, timestamps, and other relevant shot information.

Sessions follow a structured format, commencing with a warm-up period before transitioning into points, started with serves. Each player serves for six consecutive points, to ensure similar point distribution between players, while also mitigating fatigue for the serving player. Simultaneously, this setup provides sufficient continuity for players to establish comfort with their serve motion. At the time of writing this report, seven sessions have been recorded, each lasting for one hour and comprising 451 serves with their corresponding points (excluding double faults).

Data is then uploaded to a database in the form of text files with all ball and player positions as well as timestamps for the shots and other additional information.

3.3. Two model approach

As previously mentioned, the primary objective of this project is to evaluate and analyze the performance of the tennis serve. Unlike other shots in tennis, the serve does not rely on the opponent's output but rather depends solely on one's own technique and skill. Given this unique nature of the serve, it becomes crucial to establish a method for determining the quality of a player's serve accurately.

The term “ace” serves as the ultimate classification for a well-executed serve. An “ace” is a serve that ends the point immediately, leaving the opponent unable to even reach the ball, thereby awarding the point to the server outright. While aces are straightforward to identify and serve as a significant metric for evaluating the performance of professional players, they present a more complex challenge when it comes to club-level tennis.

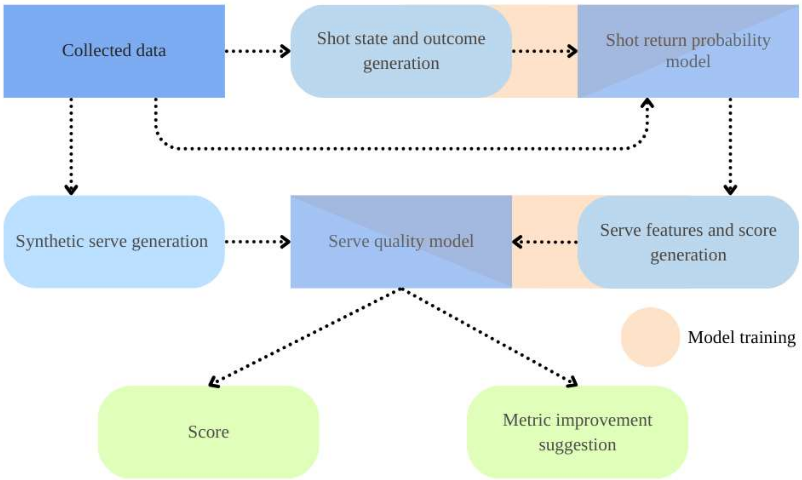
At the club level, aces are not as common, making them insufficient as the sole criterion for evaluating serve quality. This is because serves that do not result in aces can still be considered effective and valuable. Such serves can force the opponent to make a mistake or produce a weak return, thereby giving the server an advantage in the subsequent shot. Thus, it becomes essential to consider other factors and metrics to comprehensively assess the quality and effectiveness of serves beyond just the occurrence of aces.

This rationale leads us to adopt a two-model approach to analyze serve performance in greater depth:

- **Shot Return Probability Model**: The first model focuses on evaluating shot quality by examining its probability of being returned by the opponent. Unlike traditional models that assess shot quality based on isolated instances, this model provides a dynamic analysis. It calculates the probability of the shot being returned at every moment from the instant the ball is struck to the moment it bounces, offering a more comprehensive understanding of shot effectiveness over time.
- **Serve Quality Model**: As previously mentioned, a well-executed serve can provide a significant advantage in subsequent shots, influencing the overall outcome of a point. Therefore, insights gained from the Shot Return Probability Model are then utilized to inform our understanding of serve quality. By understanding how serve performance impacts the subsequent shots, it is possible to develop a qualitative classification of serve performance that takes into account its broader impact on the game. From this second model, level of importance for each feature of the serve can be obtained, so a recommendation for improving certain specific aspects of a given serve can be generated.

By employing this two-model approach, the aim is to provide multifaceted analysis of serve performance that considers both immediate and long-term outcomes. This integrated methodology allows to capture the complexities of serve quality more accurately and offers valuable insights that can assist players and coaches in optimizing their serving strategies and overall game performance.

Altohugh these two models are trained and tested separately, the Serve Quality Model feeds from data outputted by the Shot Return Probability Model, interacting with each other. The following figure shows an overview of how the whole system is developed.



2. Flow chart with the workflow of the system combining both models.

3.4. Metric and feature definition

As the project follows the two-model approach, there are two sets of features needed for each specific model. The words metric and feature are both used in similar situations throughout Section 3 and Section 4, they reference the same concept but in a slightly different context, metric is used when referring to the characteristics of a shot state or a serve, generally measured in metric system units; features refer to the metrics when used as values for the vectors imputed to train and test machine learning models, often normalized to the maximum possible value of each feature.

3.4.1. Shot state metrics.

In this scenario, the state of the point in any given moment during the trajectory of the ball between the shot impact moment and the bounce should be described and, in order to correctly predict the probability of shot return, the following metrics were chosen:

- Time since shot: time elapsed since the hitting instant to the moment when evaluated.

- Time until bounce: time remaining until bounce, in the case of the data collected, real bounce time was considered, when possibly applied to real time situations, expected bounce time would be used.
- Player to bounce distance: difference of position between bounce and player location, only in the ground 2D plane.
- Player to trajectory distance: difference of position between extended trajectory of the ball after bounce and player location, only in the ground 2D plane. Always smaller or equal to the player to bounce distance.
- Player speed to bounce: player velocity projected in the direction of the bounce (or expected bounce in live implementation).
- Ball speed: separated in the three main directions, court length (x), court width (y) and height (z), longitudinal, lateral and vertical axis respectively, as well as the absolute value of the velocity of the ball.

These metrics aim to be able to define the given state of a shot and to analyze the outcome of it.

3.4.2. Serve metrics.

Following a comprehensive investigation involving an examination of standard metrics used by coaches across different proficiency levels, coupled with an understanding of the capabilities of the PlayReplay System, different metrics were chosen to evaluate serves:

- Speed: The velocity of the ball immediately after impact, a classic metric in coaching and tournament analysis.
- Spin: The revolutions per minute of the ball post-impact, crucial for categorizing serves based on their spin type (top spin, slice, flat).
- Hitting Position: The coordinates of the ball the moment of impact, important in determining trajectory.
- Position at Bounce: The coordinates of the ball at the point of bounce, defining the subsequent trajectory.
- Height Over Net: The vertical position of the ball when crossing the net, influencing trajectory.
- Angle: The trajectory angle of the ball relative to the court's sidelines.

3.5. Data processing and shot state generation.

3.5.1. Provided data description.

After each data session, the data is uploaded and ready to process. The data generated by the system is divided in three files:

- "Shots" file: text file that contains the information of every shot that occurred during the session. The attributes include shot timestamp, shot number in a specific rally, shot type, shot outcome and top spin. Timestamp, x-coordinate, y-coordinate and z-coordinate are all part of the next features which are also included: player position, ball position at shot, ball velocity at shot, ball position at the net, ball velocity at the net, ball position at bounce and ball velocity at bounce. All the 33 attributes are represented by an integer or decimal value. Meters, meters per second and microseconds since epoch are the units chosen for describing positions, velocities and timestamps respectively.

- "Player position" file: in this text file the position of both players taking part of the session is represented with three position values, one for each coordinate, estimated as the center of gravity of the player, all values measured in meters. Alongside the player position, timestamp of each sample, once again, indicated in microseconds since epoch.

- "Ball position" file: the ball is detected and tracked during the whole session, as long as it is in play, and its three-dimensional position is calculated and provided as three coordinates' values measured in meters, paired with a timestamp for each sample.

It is worth noting the fact that sample rate for ball tracking is higher than the one used for player tracking. This explains the different values in timestamp for player and ball position at shot. It is also considered when processing the data.

3.5.2. Shot state generation.

For the purpose of this project shots are divided into several instants during the time interval between the shot hitting moment and the bounce of the ball. Each of these instants represent the state of the shot at that specific moment. The state of the shot is defined as the situation of the ball and players at a certain instant of that shot, all generated states are analyzed and used individually to train the models described in the following sections. There are two main reasons which justify the procedure:

- Limited data: machine learning algorithms and models require large quantities of data in order to give reliable and conclusive results, with the dataset used for this project more samples are generated compared to using each shot as one unique sample when training the model.
- Point state change during ball flight time: it is easy to simplify the state of a point as the state of the shot the moment its hit, but since the contact point, movement and position of the players can change the probability of different outcomes to happen, the receiving player can act in several ways, varying speed and direction to optimize the chances of returning the shot, all during a short time, in this case the period while the ball flying until it bounces.

Then, each shot needs to be divided in individual moments between the ball contact and the bounce. As described in Section 3.4.1, shot state metrics use information of both ball and player. In the data provided by the system there is a mismatch between the sample rate and detection instants of ball and player position. In other to solve this issue, only samples of player position were used, player position and speed is directly obtained and calculated from the original data, but ball information is generated.

Because of the asynchronous nature of the ball and player tracking method, cases exist where the ball is not tracked at the same exact moment as the player. Player movement is unpredictable and difficult to model, on the other hand, ball movement since the moment right after the shot can be described as a solid body movement in free space with gravity present, which can be approximated defined with a quadratic function. Ball data is then fitted to a continuous quadratic function which is applied to the player position instant, obtaining an accurate estimation of the previously unknown ball position and speed at that specific moment.

As all timestamps in the collected data are measured in microseconds since epoch, player position samples can be obtained from the “Player position” file and filtered by the time defined from the ball contact and the bounce of a specific shot, available in the “Shots” file. Ball position samples are filtered the same way but accessed through the “Ball position” file.

A total of 176 points were processed, obtaining 1970 shot state samples, with an average of approximately 11 states per shot, although it is worth noting that the range of states in a single point varies from 1 to 39.

3.5.3. Shot state metrics and outcome generation.

Each of the previously described metrics for shot states need to be generated from the raw data provided by the system after the data collection sessions. All metrics are normalized in respect to standardized maximum values, ranging them between 0 and 1, as needed when developing the logistic regression model.

- Time since shot: each shot state is processed within its corresponding shot, contact point of the ball with the racket is recorded and available in the “Shots” text file under the feature name “Ball position at shot timestamp”. Shot state timestamp is in itself the timestamp of the players position at that chosen shot state, available in the “Player position” file. Time since shot is then computed as the difference between the two, in microseconds.
- Time till bounce: similarly to the previous metric, it is calculated as the difference between the shot state timestamp and the bounce timestamp, accessible as “Ball position at bounce timestamp” feature in the “Shots” file.
- Player distance to bounce: computed as the length in meters, of the 2D vector subtraction of the bounce position, constant for each shot and therefore common for every state in the same shot, and the position of the player in that shot state instant, variable throughout the states.
- Player distance to trajectory: after the bounce, the ball continues to move until the second bounce or until the returning player hits the following shot in most cases, following a mostly straight line. A virtual straight line is drawn in the 2D plane of the court containing the points defined as the ball position at contact with the racket and the ball bounce position. Player distance to trajectory is

calculated as the minimum distance, in meters, between the player position at that shot state instant and the trajectory line, constant for the duration of the shot.

- Ball speed: the first derivative of the fitted curve computed from the samples of ball position, represents the speed of such ball during the shot. Three curves are computed, x, y and z, allowing to extract one value of speed for each coordinate at the shot state instant, measured in meters per second. Absolute ball speed in the 3-dimensional space is also obtained as the squared root of its components.

Every shot is then analyzed in the context of the point of which it corresponds, and its outcome is logged. Shots are classified into two classes, returned or not returned. If a given shot is successfully returned by the opponent, hitting back the ball into the boundaries of the court, the shot is considered returned (class 1), indicating the continuation of the point. On the other hand, if the opponent does not reach the ball before the second bounce or returns the ball outside of the limits of the court, it is classified as not returned (class 0), ending the point.

Individual data collection sessions are processed and stored for later use in the training and developing of the shot return probability model.

3.5.4. Serve metrics and score generation.

Six data collection sessions were processed to obtain 364 serves described with 9 features and a calculated score based on the shots played right after the serve. Mentioned score and features are computed the following way.

As stated before, the impact of the serve in a given point does not stop after the return of the serve. A high quality serve not always has an ace as an outcome, but if it is returned with nit such effectiveness, it creates an advantageous position for the server in the following shots by both players. Following the studied impact of serve in point by Jeff Sackmann [21], the score of collected serve is calculated. For every point, each shot is divided into its shot states, metrics for the generated shot states are obtained and inputted to the shot state model, the model then outputs a probability of the shot being returned. An average of all the shot states computed probability of return, for an individual shot, is calculated and assigned to the corresponding shot, ending with "quality" assessment of all the shots from every point. Shots are then divided into shots made by the server or the returner, and this shot "quality" given by its probability of return is then used to make a weighted average to represent serve impact on the quality of the shots. The difference between the shot quality impacted by the serve, of the two players is directly proportional to the score given to the serve of the point, a value in the range from 0 to 10, higher than 5 meaning that the average probability of return of the server shots is lower than the returner's average probability of return. Then for better visualization, a modified logistic function [22] which can be seen in Appendix C, is applied to the value, giving the final score

Similarly to the shot state metrics, serve features are obtained from the collected data, whereas in this scenario, they are all obtained from the "Shots" file, with no need for computing or calculation, except for one. For each serve, "Ball speed at shot", "Top spin", "Ball position at shot", "Ball position at bounce" and "Ball position at net" are values computed and provided by the PlayReplay system. Only the metric "Angle" needs to be calculated as the inclination of the straight line defined by the ball hitting position and the bounce position, in the court 2D plane, in relation to the longitudinal axis ("x" axis) of the court.

All serves with its corresponding features and scores are saved for future use in the serve quality model.

3.6. Machine learning models design.

In this section machine learning model selection and design are discussed and explained, from the classes on which they would be tested, to the decision to implement certain regression and/or classification models for the cases of this thesis.

3.6.1. Shot Return Probability Model

First implemented model is the shot return probability model. The goal of this model is to, given a state of a shot defined by the previously described metrics, at any point during the flight time of the ball between the contact moment and the bounce; estimate the probability of that shot to be returned.

For this task, a logistic regression model is chosen. Logistic regression models are often used to do binary classification by providing a probability for a given sample to belong to one class out of two. In this case the two classes are: shot returned and not returned. It is needed to mention that the model is not used to classify a shot and therefore predicting the outcome, its goal is to give a probability of a shot being returned, independently of the real outcome of the shot. This probability is understood as the "quality" or "decisiveness" of the shot, the more likely it is to be returned, the less decisive it is, on the other hand, the less probable to be returned, the higher the chances to win the point by the player hitting the shot, immediately after or in the following shots.

The data used to implement the model is formed by 1970 shot states defined by the mentioned 9 features normalized to values between 0 and 1. The classes are represented with 1 when the shot for which that shot state is part of is returned and 0 when it is not returned. The data is split into training and testing sets to visualize the performance of the model in Section 4.1.

An unmodified logistic regression model [23] from the sci-kit learn python package is trained, tested and saved to be used in later stages of the project.

3.6.2. Serve quality Model.

In this case gradient boost regression is the method used for this task. A total of 364 serves, six data collection sessions, each described as the values of its 9 features or metrics and tagged with a previously calculated score based on the shots made after the serve, as described in Section 3.5.4. A gradient boost regression model [24] is acquired from the sci-kit learn package, modifying the number of estimators parameter to 18, to improve performance given the used data size.

Considering the limited number of serves used to train the model, maximizing training data is key to improve the performance of the model, for such reason, leave one out cross validation technique [25] is used to show and analyze performance. The leave one out technique consists of iterating an amount of training and testing data splits equal to the number of samples in the data, for every iteration, only one sample is used for testing allowing the rest to train the model. Performance of the model is then seen as the combined performance of all iterated models in all samples when being part of the testing set.

Sensitivity analysis is performed to measured how changes in different features can affect the outputted value of the score, to be shown in Section 4.3.2.

To illustrate the possible use and analyze the performance of the model in a real live scenario, a set of virtually generated serves are computed, each one with a plausible and unique combination of values to simulate a possible serve training session. The mean of every feature is calculated to generate an average serve for the session, which is then inputted to the model obtaining a score and most importantly a set of feature sensitivity for that specific average serve.

3.7. Project development workflow.

Even though the structure of this methodology is constructed for a better understanding of the processes, technics and overall final system, as described in Figure 2, during the development of this project the taken steps follow this order:

1. Collect the data.
2. Generate shot states, shot states features and outcomes.
3. Develop, train and test the Shot Return Probability Model.
4. Generate serve features and scores.
5. Develop, train and test the Serve Quality Model.
6. Generate synthetic serve data.
7. Simulate a digital coaching example.

4. Results

4.1. Shot return probability model results.

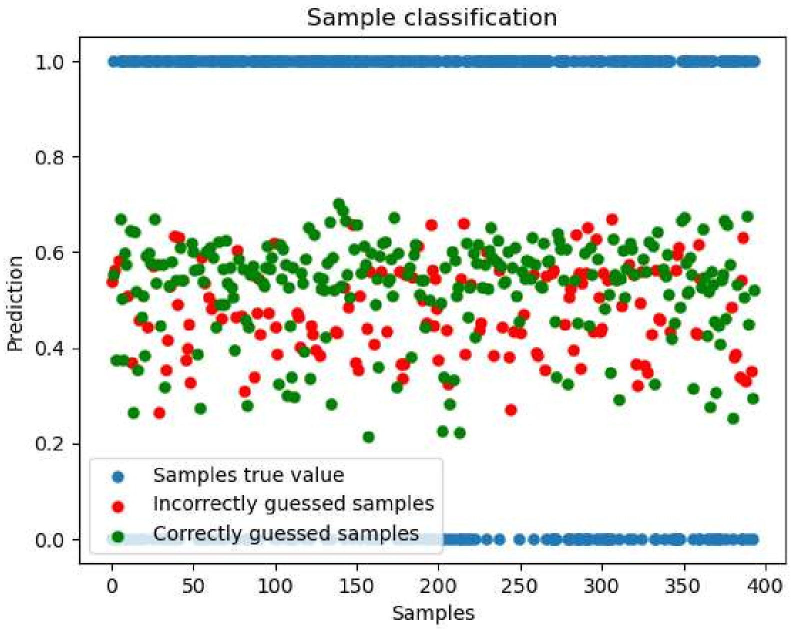
From the collected data, one one-hour session, 1970 samples of shot states were extracted and used to train and evaluate a Logistic Regressor model.

Model was trained on samples being labeled as 1, meaning the shot was returned; and 0, shot not returned. The predicted value outputted by the model would mean the probability a given sample to end up being returned.

Although in the following analysis of performance classification of shots is used to assess the accuracy of the model, it is important to clarify that the goal of the model is not to classify or predict the outcome of a shot, but give a value to estimate the "decisiveness" of the shot, tennis is a highly unpredictable sport and classifying shot predictions as returnable or not returnable could, in many cases, not be the correct way to approach analysis of the game. Having said that, it sure can be useful to interpret the capabilities of the players and understanding the sport dynamics.

4.1.1. Accuracy on the testing data.

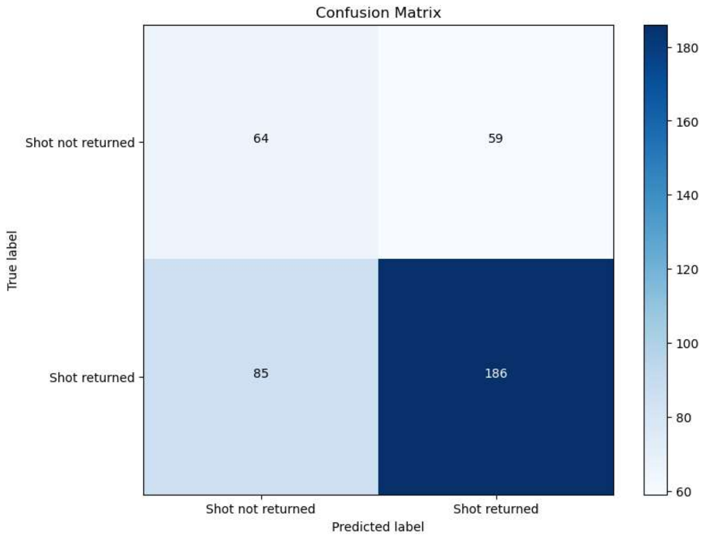
The developed model was able to achieve the following results on the testing data:



3. Sample classification on the testing data: in blue, the true value of the outcome of the sample; in green, shots correctly classified as returned or not returned (threshold = 0.5); in red, shot incorrectly classified.

The previous image shows predicted probabilities of return for the testing samples, classified as correct or incorrect based on thresholding. Shots with a lower value than 0.5 are classified as not returned and shots with a higher or equal value than 0.5 are classified as returned. Image shows the correct and incorrect guesses.

The accuracy and confusion matrix obtain from this data is the following:



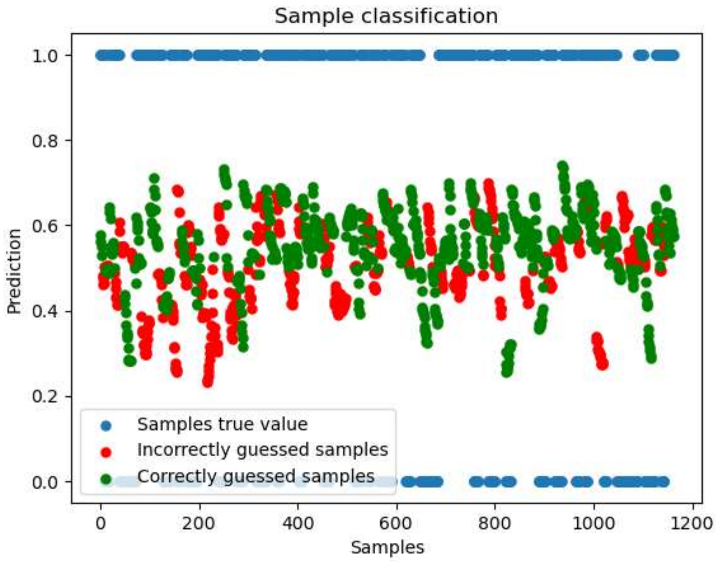
4. Confusion matrix for the testing data, shot return probability model.

Accuracy is calculated to be 0.634 for the given test set. There should be taken into consideration how this model aims to output the probability of a shot being returned and not to classify the shots, the accuracy value of the model gives an approximate idea of the performance and consistency of the players.

4.1.2. Accuracy on an external set of samples.

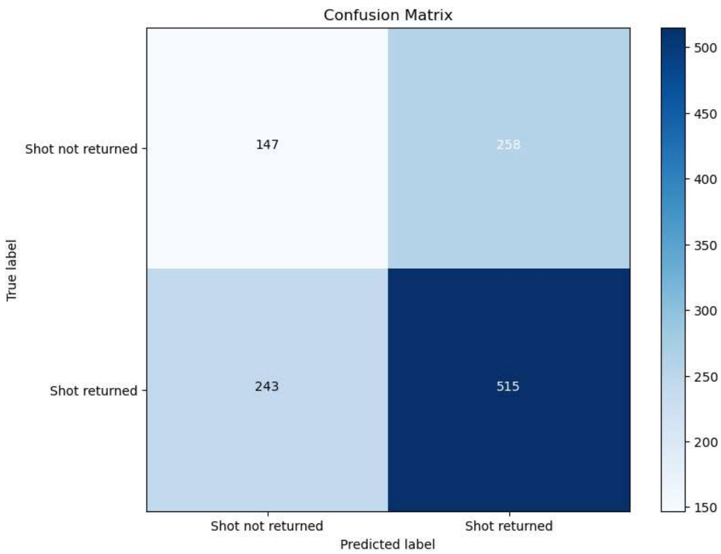
Different samples can be part of the same shot, for that reason, and due to the limiting data, it is reasonable to think that the model would have good performance in classifying samples from the test data that are similar to others used in the training data.

In this case the model is evaluated with 898 samples extracted from shots not used to train the model.



5. Sample classification on external data: in blue, the true value of the outcome of the sample; in green, shots correctly classified as returned or not returned (threshold = 0.5); in red, shot incorrectly classified.

In this case, the model seems to be less precise when classifying the shots, giving the following confusion matrix:



6 Confusion matrix for external data, shot return probability model.

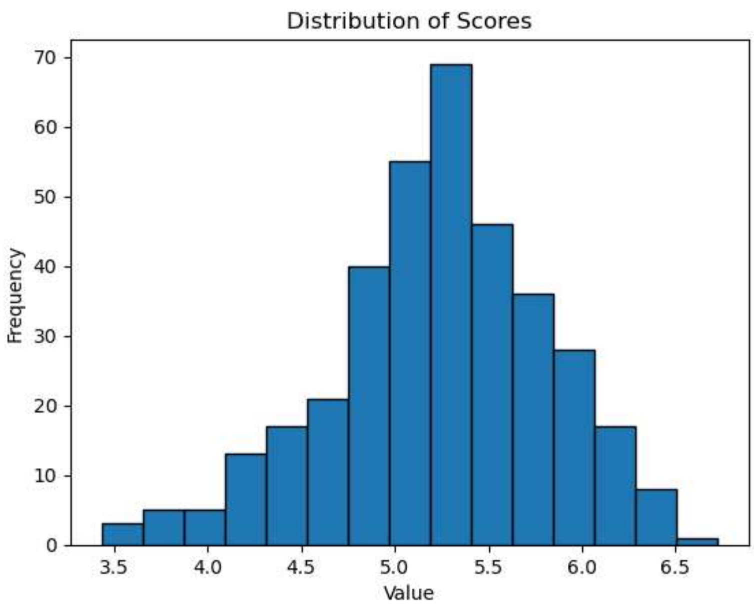
Accuracy lays to 0.569, similar value as the testing data set. Further discussion regarding these results follows in Section 5. This decrease in accuracy can be explained by the use of a different set of points in relation to the training data, shot states used in this performance analysis belong to shots not present in the training set, which although more representative of a real world scenario, the lack of multiple shot scenarios in the current dataset might give lack of precision for certain combination of values for a shot state.

4.1.3. Serves scoring.

After the model is trained and tested it is used to obtain probabilities of a shot to be returned for every shot in six data collection sessions. As explained in Section 3.5.4, serve score is calculated as combination of shot return probabilities from both players during the point, meaning that serve score is highly dependent on return probability model performance.

It is then interesting to take a closer look at the obtained serve scores and their relationship with the serve features.

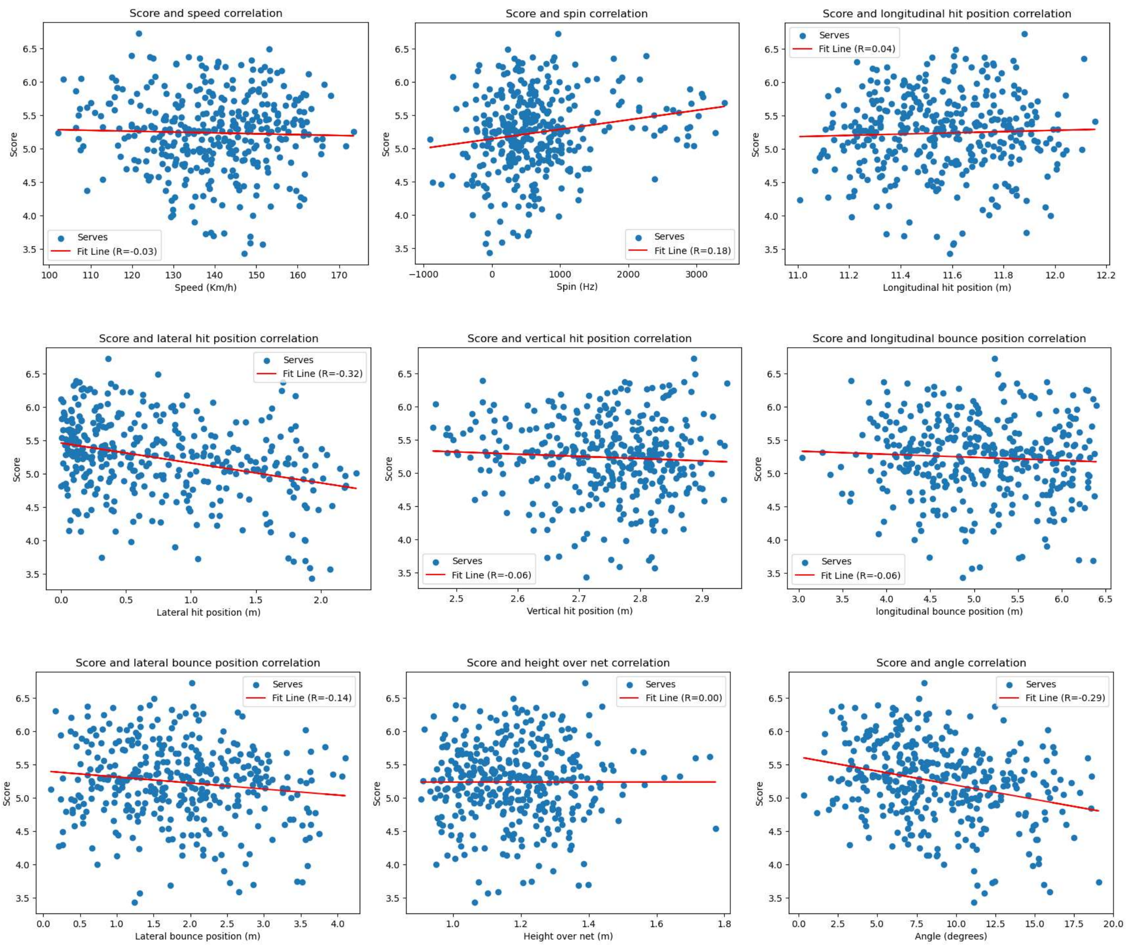
The 364 serves have the following distribution of scores:



7. Histogram of scores.

The figure shows an approximate normal distribution of scores with a peak frequency between the score values 5 and 5.5, with an average of 5.23, suggesting that even not at professional level, serves do indeed provide a slight advantage during points, as a value higher than 5 shows that the shots hit by the server during the point. In addition, 252 serves or 69% of the serves have a score higher than 5, supporting the idea of serves being relevant and important to improve, as it provides a higher probability of winning points.

How serve score is related to different features can give an understanding of which of those metrics are more important when producing a high-quality serve.



8. Correlation between calculated score and serve features. Serves represented as blue dots and correlation or fit line in red.

The previous figure shows how each feature is correlated with the score of the serves obtained based on the "quality" of the following shots in the point. A low correlation between most features of serves can be observed, although there exist interesting details: both lateral hitting position and angle have negative correlation to score with means that shots hit close to the center of the court and aimed to bounce on the "T" side, side of the serve box further from the lateral limit of the court, tend to obtain better score.

The low correlation can be explained by different arguments, firstly, the data set is limited, not only in the number of serves, but in relation to the variety of conditions on which they were captured, diverse players, court surfaces, and contexts are not present in this data set, more to discuss in Section 5.4. In addition, features can be dependent on each other, the speed of the serve for example is limited by the angle of which it is hit, the net is higher at the sides of the court, so the broader the angle the more curved the trajectory of the ball must be, in order to clear the net and land on inside the serve box. Analyzing individual metrics by themselves can be useful but fails to interpret the complexity of the problem.

4.2. Serve quality model results.

4.2.1. Model performance analysis.

As mentioned, the leave one out cross validation technique is used in this case to maximize use of available data to analyze model performance.

For each iteration where a new model is trained, only one sample is tested, predicted and actual values of the serve score are stored and compared to determine the performance of the model.

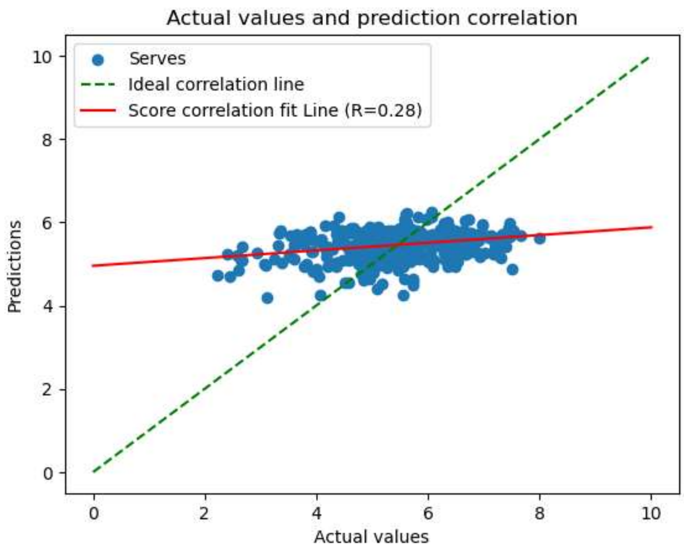
A regression model can be analyzed with two parameters: mean squared error and R^2 score or coefficient of determination.

The mean squared error is the average squared difference between the true value, in this case of a serve score, and its prediction, meaning that the smaller the mean squared error, the better the performance of the model, as the predictions are close to the actual values.

The R^2 score, lower than 1, provides a measure of how well the predicted values from the model match the actual values, in this case the score, the higher the value, the more accurately the model predicts the serve score. The coefficient of determination can be lower than 0 meaning that the model does a poor job when predicting the scores.

When using the full data set of serves available the values for the two described parameters are: 1.11 mean squared error and 0.07 of R^2. These values can indicate a not precise prediction capability of the model, which can be explained by the size of the data set and the level of inconsistency of the players.

To analyze deeply the accuracy of the regression, predicted values can be compared and plotted directly against the actual values of score.



9. Correlation plot between actual score values and predicted values.

As seen in the figure, there seems to be a small correlation between predictions and actual values, showing an increasing tendency in predictions for higher scored serves. Nevertheless, as the correlation coefficient shows, with a value of 0.28, predictions differ significantly to actual values.

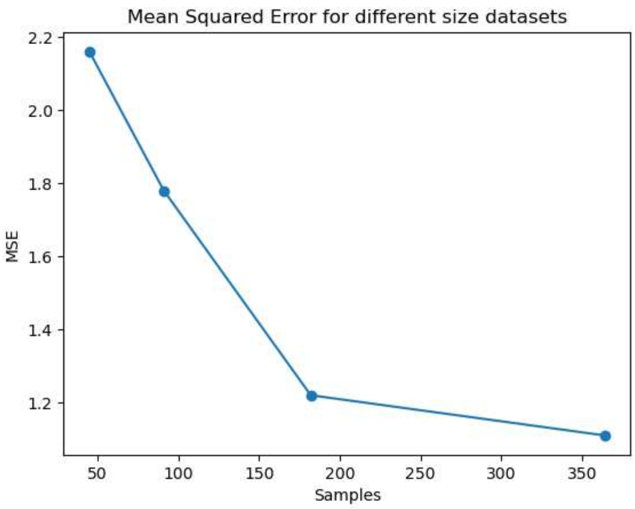
4.2.2. Dataset size impact.

One of the reasons mentioned for this reason is the size of the data set used to train the model. The relationships between feature values and scores cannot be assumed to be linear and it is reasonable to believe that the score values given by the probability of the shots being return during the points is noisy, considering the conditions for the shot return probability model, explained further in following sections. For these reasons the size of the dataset seems to be an important factor for the performance of the model.

By observing the performance of the model with different sized datasets, extracted from the current group of scored serves, ideas of performance for larger, currently nonexistent datasets can be inferred.

Four different models were trained, each one with double the amount of serve samples then the previous, the last one using all available samples. Performance parameters for mentioned models are calculated.

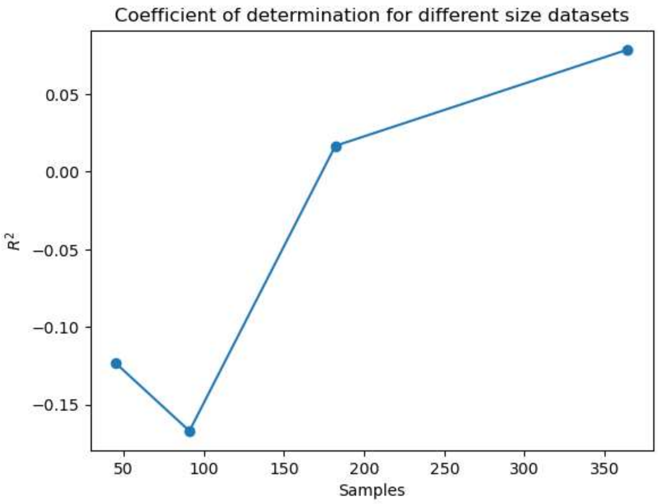
The mean squared error progresses as follows:



10. Mean squared error for serve quality model trained with different size datasets.

As seen in the figure, the mean squared error of the model seems to decrease considerably the more samples are used to train the model, as expected, in this case, the downwards tendency does not seem to disappear yet when reaching the biggest dataset of 364 serves.

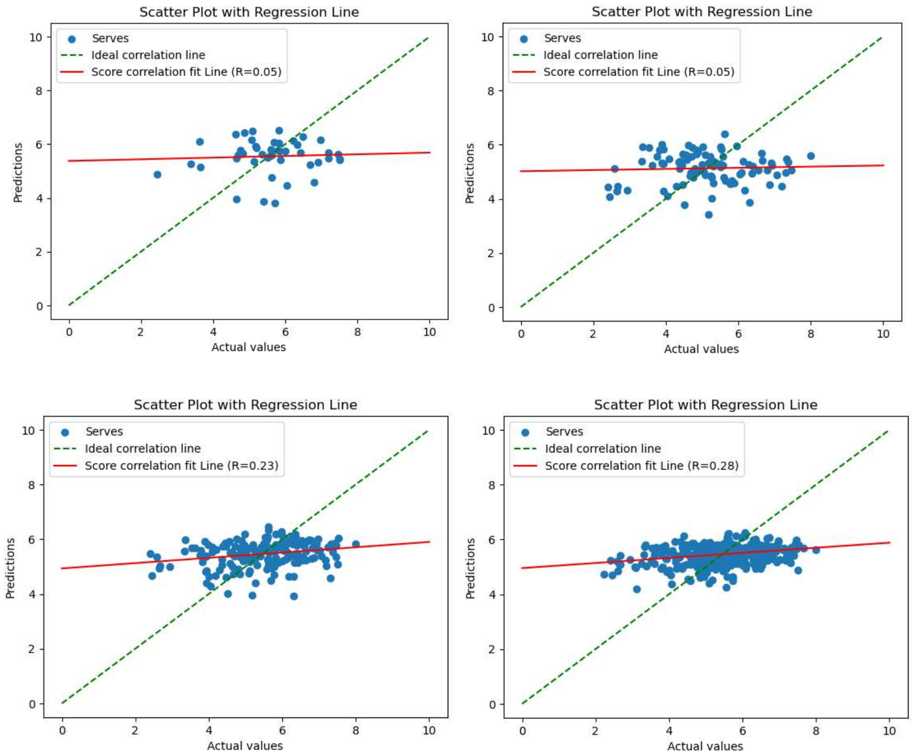
In the case of the coefficient of determination, the result lays as shown:



11. Coefficient of determination for serve quality model trained with different size datasets.

Similarly to the previous parameter, R^2 shows change when increasing the number of samples, in this case an uptrend can be seen, again not reaching a visible limit or plateau, suggesting a possible increase in performance when using larger datasets than 364 samples.

Actual and predicted values can as well be calculated and plotted for each model:



12. Correlation plots between actual score values and predicted values for serve quality models trained with different sized datasets.

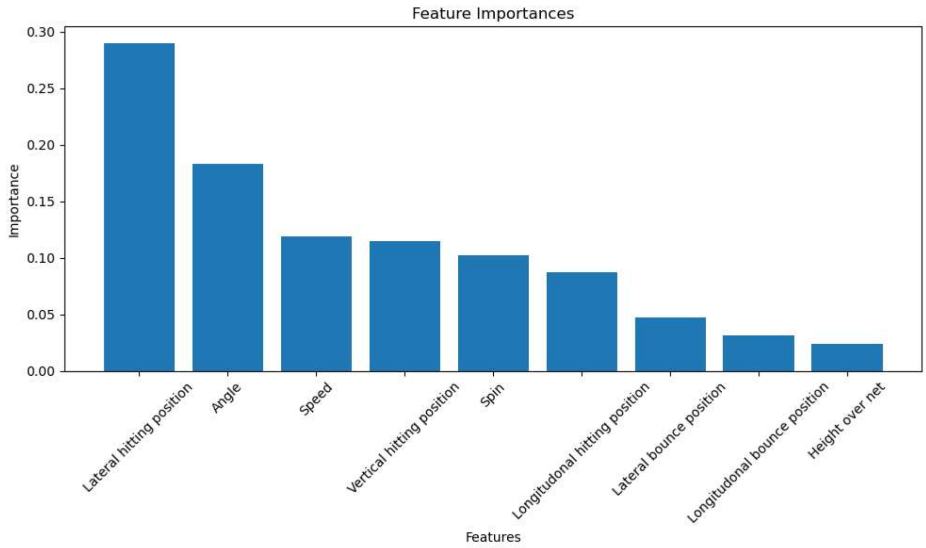
Correlation also increases as more samples are used, supporting the idea of better performance achievable with a bigger dataset.

4.3. Serve features importances.

One of the purposes of the serve analysis proposed in this project is the idea of studying the contribution of the different serve metrics to the overall score of the serve, with gradient boost regression feature importance can be obtained but in addition to that, the goal of the digital coaching is to provide an understanding of which metric change can contribute to an improvement in serve performance.

4.3.1. General feature importances.

When training the model with the 364 serves available, feature importances are ranked as shown:



13. Feature importances ranked from most important to least.

Lateral hitting position appears as the main metric to determine the score of the serve, being one of the metrics with the highest correlation to score, shown in Section 4.1.3. Angle follows as the second most important metric, also with high relative correlation to score. Importances of the following metrics considerably decrease, height over net being the least contributing feature.

4.3.2. Feature importance for a given serve: digital coaching example.

Although general feature importances can provide insights about how metrics influence serve quality on average, numerous serve techniques and tactics exist, providing an immense range of possibilities and combination of metric values. It is interesting then to study the feature contribution to specific serves and how the score changes when only one metric is modified.

This analysis amplifies the utility of this serve quality model, as it is immensely difficult to study the change of one metric when leaving the rest constant by recording enough points for each metric variation to have a statistically reliable result. Time, effort, fatigue and consistency are limiting factor to this idea, even professional level players could not maintain consistent serves only modifying one metric during the recording of multiples serves.

For this example, a set of 30 virtual serves is generated by modifying the metric values of real recorded serves. Each metric is increased or decreased randomly by 0 to 10% of the maximum possible value of the metric. This virtual serve set tires to mimic a possible serve training session, where a player hits 30 serves to analyze in order to understand their average performance.

For the given set of virtual serves an average value of each feature is computed to generate an "average serve". The serve quality model is then used to give a score to the average serve and sensitivity analysis of the model is made in the context of the given serve.

An example of a virtual serve:

Serve Metrics				
Speed (km/h)	Spin (Hz)	Longitudinal hitting position (m)	Lateral hitting position (m)	Vertical hitting position (m)
136.49	671.76	11.50	0.79	2.73
Longitudinal bounce position (m)	Lateral bounce position (m)	Height over net (m)	Angle (°)	Score
5.21	1.75	1.25	17.26	5.76

14. Simulated serve metrics values.

Then sensitivity analysis is performed on the given serve. Sensitivity analysis helps determine how changes in individual serve features impact the overall score predicted by the model. By systematically varying each feature while keeping others constant, sensitivity analysis identifies which features have the most significant influence on the score. This allows you to understand which aspects of a serve are most critical to achieving a higher score, enabling targeted improvements in serve performance. This analysis is particularly useful for optimizing serves, as it reveals how tweaking specific features can lead to better outcomes.

When applying a 5% variance of the maximum possible value for each feature, the following score changes are obtained:

Metric sensitivity				
Speed	Spin	Longitudinal hitting position	Lateral hitting position	Vertical hitting position
0.63	1.55	0.58	-2.24	0.41
Longitudinal bounce position	Lateral bounce position	Height over net	Angle	Maximum score
0	0	0	-0.27	7.31

15. Metric sensitivity for the simulated serve.

Meaning that if the given serve is modified by one of the metrics by 5% of its maximum possible value it would change its score by the amount shown. For instance, if the speed is increased 14.47 km/h, to be then 150.96 km/h, the score of the new serve is 0.63 higher. In this example, the metric change that increases the serve score the most is the spin, bringing the score 1.55 points higher to 7.31.

This approach helps to understand serve performance on a deeper level and gives concrete information about what aspect of the serve is more important to improve by the player in order to get better results.

5. Discussion

The objective of this project was to develop a reliable approach for evaluating the quality of a tennis serve through the application of a data-driven methodology. The utilization of a machine learning model served as a tool for quantifying and analyzing the probability of shot return and the effectiveness of various serve characteristics. The core of this discussion is based on the insights derived from the analysis, their implications for tennis performance, and the broader significance these findings hold within the context of sports technology.

5.1. Interpretation of Results

The outcomes of this project reveal several key factors that contribute significantly to a successful tennis serve. Among the nine features evaluated, serve speed, placement, and spin emerged as particularly critical to general serve performance. These findings align with established tennis theories, which underscore the importance of these elements in various types of tactics.

Additionally, the sensitivity analysis conducted on the model gives insight about the relative importance of each feature for a given serve. While certain characteristics are generally significant, their impact may vary depending on the context, such as the player's style, tactic chosen, first or second serve, etc. This nuanced method shows the importance of a comprehensive and personalized approach to serve training, where multiple aspects of the serve are individually analyzed to help individuals understand their weaknesses and strengths.

Although used as a tool to generate serve score, the shot return probability model stands as a novel implementation of technology to assess quality of shot performance. Classic metrics for tennis performance analysis such as unforced errors classification are currently not consistent and subjective to the calculation criteria of the statistics generators. This example of objective analysis sets an example for future implementation of statistics such as forced or unforced errors.

5.2. Implications for Tennis Training and Performance

The findings from this project carry significant implications for tennis coaching and player development. By quantifying shot and serve quality, coaches can focus specific areas where a player excels or requires improvement. This objective analysis complements the subjective evaluations traditionally used in coaching, offering a more comprehensive assessment of a player's performance in different scenarios. For instance, if a player consistently performs high serve speeds but lacks accuracy, training can be adjusted to improve precision without compromising power.

Moreover, the detailed analysis of serves opens up new ways for match preparation and strategy development. Players can leverage this data to train and improve serves that exploit their opponents' weaknesses or perfect their techniques based on the specific conditions of the match, such as court surface or weather. This strategic application of data can provide players with a competitive advantage, particularly at higher levels of play, where the margin for error is often critical.

5.3. Broader Significance in Sports Technology.

Beyond its immediate application to tennis, this study contributes to the broader field of sports technology by exemplifying the value of data-driven analysis in performance evaluation. The increasing availability of detailed data has made quantitative methods an essential tool for assessing athletic performance. This study demonstrates how such data can be harnessed to gain deeper insights into the mechanics of sports technology, a methodology that can be applied across a multitude of disciplines.

Specifically, the approach adopted in this study could be adapted to other sports where similar principles apply, such as volleyball serves, soccer penalties, or basketball free throws. The ability to quantify and analyze performance in these contexts could lead to the development of more effective training programs and, ultimately, better overall performance.

5.4. Limitations and Areas for Future Research

While the findings of this project are promising, several limitations should be considered. The relatively small dataset, though managed effectively through cross-validation, restricts the generalizability of the results. A larger dataset could support a more robust analysis for both the shot return probability model and the serve quality model. More in depth serve technique features, coming from skeleton tracking, could as well provide more information on serve quality.

Not only the size but the variety and complexity of the dataset. Only four male right-handed players with a similar playing level participated in the data collection, which makes the developed models susceptible to poor performance when analyzing shots and serves from player with different characteristics to the ones described. Future research should focus on first separating different group of players and developing several models trained just with data form specific demographic of athletes so these models can be sed in distinct situations, each model trained with sufficient amount of data for each athlete group.

Moreover, this project primarily focused on the technical aspects of the serve, without considering other influential factors such as psychological pressure, fatigue and environmental conditions. These factors can significantly impact serve performance and should be incorporated into future research. For example, analyzing serves under different levels of psychological stress could offer insights into how players adjust their techniques in high-pressure situations.

Future research could also explore the use of more sophisticated models capable of capturing non-linear relationships between serve features. While the current model serves as a solid foundation, more advanced techniques could yield even more accurate predictions and provide deeper insights into the determinants of serve quality.

6. Conclusions

This project has made a contribution to understanding tennis serve quality by developing a method for quantifying and analyzing the key features of a successful serve. It is demonstrated that a combination of factors—such as speed, placement, and spin and others—are crucial to serve performance, and that these factors can be objectively measured and enhanced through targeted training.

The implications of these findings extend beyond tennis, offering valuable insights for the broader field of sports technology. The approach taken in this study can serve as an example for other sports, where performance can be similarly quantified and analyzed to drive improvements in training and competition.

However, this study also highlights the importance of considering a wide variety of factors, both technical and non-technical, in evaluating athletic performance. The limitations identified suggest that there is still more to learn about the complexities of sports performance, and future research will need to address these challenges to build on the foundation described by this study.

In conclusion, this thesis represents a meaningful advancement in applying data-driven analysis to sports performance. The findings provide valuable insights for coaches, players, and researchers, and underscore the potential of quantitative methods to enhance the understanding of athletic performance. As data continues to play an increasingly integral role in sports science, projects like this are helpful in shaping the future of training and competition.

7. References

[1] Hawk Eye Innovations, “Hawk Eye.” Accessed: Dec. 15, 2023. [Online]. Available: https://www.hawkeyeinnovations.com/

[2] PlayReplay AB, “PlayReplay AB.” Accessed: Dec. 15, 2023. [Online]. Available: https://www.playreplay.io/

[3] G. Gerdin, P. G. Fahlström, M. Glemne, and S. Linnér, “Swedish Tennis Coaches’ Everyday Practices for Creating Athlete Development Environments,” *Int J Environ Res Public Health*, vol. 17, no. 12, 2020, doi: 10.3390/ijerph17124580.

[4] Tennis Data Innovations, “Tennis Data Innovations.” Accessed: Dec. 15, 2023. [Online]. Available: http://tennisdata.com

[5] Y. Fetisova, J. Zois, M. Spittle, and A. Dawson, “Coaches’ internal model of the tennis serve technique: Knowing or understanding?,” *Int J Sports Sci Coach*, vol. 16, no. 3, pp. 568–584, Jun. 2021, doi: 10.1177/1747954120984879.

[6] E. Roanes-Lozano, E. A. Casella, F. Sánchez, and A. Hernando, “Diagnosis in tennis serving technique,” *Algorithms*, vol. 13, no. 5, May 2020, doi: 10.3390/A13050106.

[7] O. Kolbinger and M. Lames, “Ball trajectories in tennis-Lateral and vertical placement of right handed men’s singles serves,” *Int J Perform Anal Sport*, vol. 13, no. 3, pp. 750–758, 2013, doi: 10.1080/24748668.2013.11868686.

[8] H. Gillmeister, *Tennis: A Cultural History.* Leicester University Press, 1997. [Online]. Available: https://books.google.se/books?id=AACDAAAAMAAJ

[9] "TENNIS | English meaning - Cambridge Dictionary." Accessed: Aug. 12, 2024. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/tennis

[10] International Tennis Federation, "ITF RULES OF TENNIS RULES OF TENNIS CONTENTS," 2024.

[11] T. Horvat and J. Job, "The use of machine learning in sport outcome prediction: A review," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 5, p. e1380, 2020, doi: https://doi.org/10.1002/widm.1380.

[12] S. Hanna and Jonathan H. Chan., "Machine Learning Applications to Sports Injury: A Review," in *Proceedings of the 9th International Conference on Sport Sciences Research and Technology Support - icSPORTS*, SciTePress, 2021, pp. 157–168.

[13] N. Chmait and H. Westerbeek, "Artificial Intelligence and Machine Learning in Sport Research: An Introduction for Non-data Scientists," *Front Sports Act Living*, vol. 3, 2021, doi: 10.3389/fspor.2021.682287.

[14] "About Catapult | Sport Technology." Accessed: Aug. 12, 2024. [Online]. Available: https://www.catapult.com/company/about-catapult

[15] "What Is Machine Learning (ML)? | IBM." Accessed: Aug. 12, 2024. [Online]. Available: https://www.ibm.com/topics/machine-learning

[16] J.-W. Kim, J. Namgung, Y.-S. Moon, and M.-J. Choi, "Experimental Comparison of Machine Learning Models in Malware Packing Detection," in *2020 21st Asia-Pacific Network Operations and Management Symposium (APNOMS)*, 2020, pp. 377–380. doi: 10.23919/APNOMS50412.2020.9237007.

[17] Z. Liu and H. Chen, "A predictive performance comparison of machine learning models for judicial cases," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–6. doi: 10.1109/SSCI.2017.8285436.

[18] D. W. Hosmer and S. Lemeshow, "Introduction to the Logistic Regression Model," in *Applied Logistic Regression*, 2000, pp. 1–30. doi: https://doi.org/10.1002/0471722146.ch1.

[19] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[20] “Tennis Abstract: Match Charting Project Metadata.” Accessed: Apr. 17, 2024. [Online]. Available: https://www.tennisabstract.com/charting/meta.html

[21] “Measuring the Impact of the Serve in Men’s Tennis - Heavy Topspin.” Accessed: Aug. 25, 2024. [Online]. Available: https://www.tennisabstract.com/blog/2017/09/06/measuring-the-impact-of-the-serve-in-mens-tennis/

[22] M. Artzrouni, “Mathematical Demography,” *Encyclopedia of Social Measurement, Three-Volume Set*, vol. 2, pp. 641–651, Jan. 2005, doi: 10.1016/B0-12-369398-5/00360-1.

[23] “LogisticRegression — scikit-learn 1.5.1 documentation.” Accessed: Aug. 26, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[24] “GradientBoostingRegressor — scikit-learn 1.5.1 documentation.” Accessed: Aug. 26, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

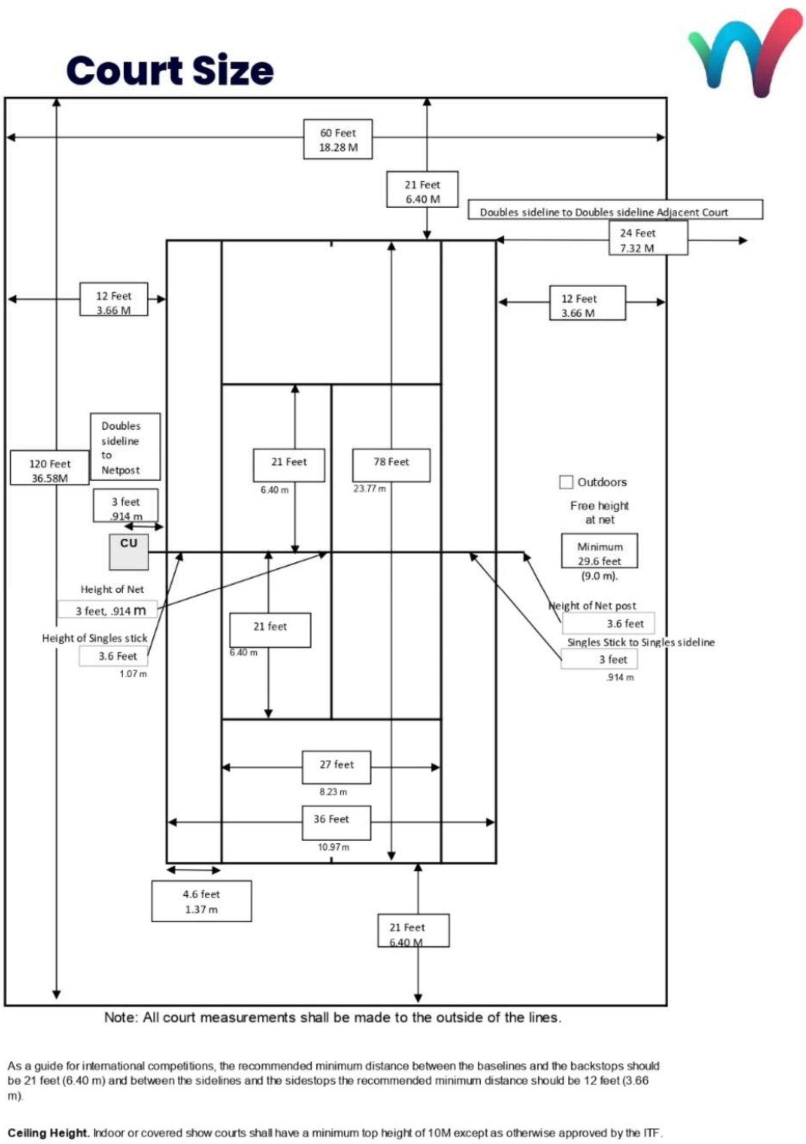
[25] Y. Zhang and Y. Yang, “Cross-validation for selecting a model selection procedure,” *J Econom*, vol. 187, no. 1, pp. 95–112, Jul. 2015, doi: 10.1016/J.JECONOM.2015.02.006.

[26] “International Tennis Federation | ITF.” Accessed: Aug. 26, 2024. [Online]. Available: https://www.itftennis.com/en/

8. Appendixes

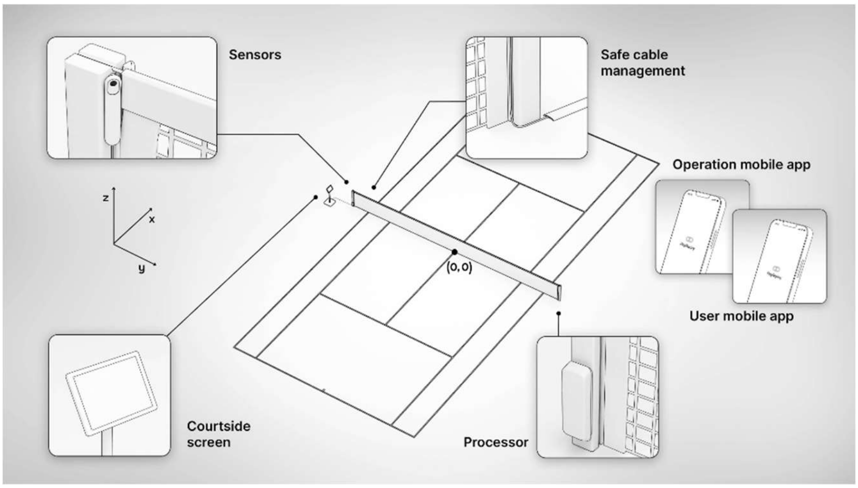
A. Tennis court dimensions.

The International Tennis Federation [26] defines the tennis court as shown:



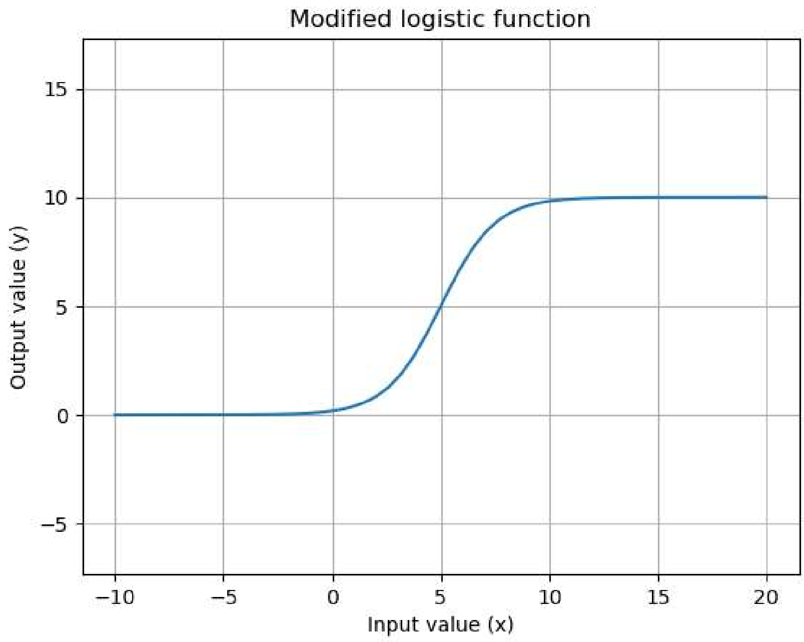
16. International Tennis Federation size definition of a tennis court.

B. Court layout and coordinate system.



17. Court layout and coordinate system.

C. Modified logistic function for score calculation.



18. Modified logistic function.

$$y = \frac{1}{1 + (e^{(-0.8(x-5))})} * 10$$