Performance Analysis and Prediction in Grassroots Football: The Use of GPS Analytics, Machine Learning, and Deep Learning

Stanley Osondu, Alsmadi Hiba

School of Computing, Engineering and Digital Technologies, Teesside University

DOI: https://doi.org/10.51584/IJRIAS.2025.10060012

Received: 19 May 2025; Accepted: 26 May 2025; Published: 28 June 2025

ABSTRACT

This study examines the application of machine learning and deep learning techniques for performance monitoring and prediction in grassroots football. Using GPS tracking data collected over an entire season, we analyze player movements, heatmaps and high-speed running activities during training and competitive matches. The research focuses on two playing positions: Central Midfielder and Left Wing. We implement six machine learning models to predict player performance and compare their accuracies. Our findings reveal significant differences in physical demands between match and training sessions across playing positions. The study demonstrates the potential of data analytics in informing player development, detecting injury risks, and enhancing decision-making in grassroots football.

Keywords: Machine Learning, Deep Learning, GPS Analytics, Football Performance, Grassroots Football

INTRODUCTION

Adopting technologies to improve sports performance has become increasingly prevalent, particularly in football. Machine learning and deep learning techniques can now be applied to large datasets collected from football players to analyze and gain insights that can enhance team and player performance. This study uses GPS-tracked player performance data to understand the differences between training and match sessions, considering how this can help transition individual players from the grassroots level to professionalism.

Our research utilizes historical player data collected using GPS devices to profile different metrics observed during training sessions and competitive matches. The data consists of approximately fifty data points collected from players in a grassroots football club, FC Pudsey in Leeds, United Kingdom, across the 2022/2023 season.

The primary objectives of this study are:

1. To determine the physical demands of training and match sessions for different playing positions.
2. To implement and compare machine learning and deep learning algorithms for performance prediction.
3. To demonstrate the potential of data analytics in informing player development and injury risk detection.

To cross-check this, the current study has implemented and compared machine learning and deep learning algorithms. Also, proposed statistical techniques, to predict the performance of the player using the current data and develop a model that evaluates the player based on their performance across all aspects of play. In the study, six regression models namely Linear Regression, Random Forest Regression, Decision Tree Regression, Ridge Regression, Gradient Boost Regression, Support Vector Regression, and deep learning model Artificial Neural Network are implemented. From the results which are carried out independently for different playing positions, it can be observed that Decision Tree and Random Forest models outperformed any other machine learning models and predicted the player's performance more accurately compared to other regression models applied.

LITERATURE REVIEW

Bradley et al. discussed the changes in different positions of play by high-intensity running patterns of elite soccer players in the English Premier League [1]. Barnes et al. studied different physical demands on premier league football players which have a seasonal factor, highlighting the increase in the number of sprints and sprint distances between 2006 and 2013 [2]. Bush et al. discussed the physical performance parameters have been influenced due to position-specific evolution in over seven seasons of Premier League football and argued that the development of tactics has played a role in the physical demands at different positions [3]. In conclusion, there is a sizable body of research on physical activity needs that has been examined through tracking data of professional soccer players during competitive matches, with a focus on positional variations and high-speed running activities. However, only a little analysis has been done on football training demands or any connection between competitive games and training schedules. Furthermore, it is yet to be established that there is any optimal model capable of manipulating training loads to achieve peak match performance. Many authors propose future work to analyze performance tracking data during competitive games to develop specific training programs [1,3,4]. There may be a possibility of constructing scientifically precise training programs that combine all training aspects, including technical, tactical, and physical if patterns between training and match needs are identified.

Verhelst et al., [1] have also proposed a linear regression model and developed the model for overall performance and for the market value evaluation of the player. Wolfgang Potthast [2] proposed a framework for an MCDM (multicriteria decision making) where they collect the data from (1) vertical jump tests, (2) 30m speed tests (4) YOYO tests (5)10m shuttle runs (6)Hoff test the data is collected from these tests and the model gives the rating of the players. In modern football, not only player evaluation can be done but also prediction and implementation of the outcome of the game using a machine learning algorithm by Sixto et al., [6] using LR, RF, ANN, and SVM, NB by Richard Pariath [8], Training and testing are done by cross-validation comparing all the algorithms. Nilay Zaveri et al., [5] developed a machine learning algorithm of the data obtained from transfermarket.com and compared the result with the original data; this way, certain errors were overcome in the already available data to evaluate the player correctly, and EVP was developed by Ihsan [6]. The EPV model is based on an output-developed CCR application. This player evaluation, like the sprinting of players proposed by Gerrard B. [7], can be used to scout players. They created a dashboard in Tableau to aid in this comparison. From the literature survey, it can be observed that the football player performance prediction model uses a small amount of data. It was also observed that various machine-learning models are built using algorithms like regression, SVM, random forest, etc. The proposed model uses large amounts of data along with neural network algorithms for performance prediction.

Petersen et al. (2018) discuss the use of technology in sports performance analysis. The authors give a general review of the many technologies, including video analysis, GPS tracking, and wearable sensors, which are frequently employed in sports performance monitoring. They also include instances of how these technologies have been employed to enhance athletic performance as well as a discussion of the advantages and restrictions of each technology.

A review article by Liu et al. (2019) focuses on the use of data analysis techniques in sports performance analysis. The authors offer a summary of the various kinds of data that can be gathered in sports, including physiological data, biomechanical data, and performance data. Also, they go over several methods for analyzing data, including machine learning, statistical analysis, and data visualization. An overview of the many modeling approaches frequently used in sports performance analysis is given by Rabitti et al. (2020). The authors talk about how mathematical models can be used to assess several facets of athletic performance, including tactical analysis, injury prevention, and athlete tracking. Additionally, they give instances of how these models have been applied to various sports, including cycling, basketball, and soccer.

METHODOLOGY

Data Collection

Data was collected using a wearable GPS monitoring device (Model: PD-99) from PlayerData Scotland. The device recorded player movements throughout all training and game days over a full season (August 2022 - April

2023). Two distinct playing positions were selected for analysis: Central Midfielder (CM/CDM) and Left-Wing (LW).

For the sake of this research, objective data was collected and analyzed from two players in an open age Grassroot football team over a season, spanning eight months (August 2022 – April 2023). For the analysis, two distinct play positions are selected: Central Midfielder (CM/CDM) and Left-Wing (LW). This is independent data directly collected using the GPS tracking device in line with the ethical demand's approval. It contains 14 Variables and 50 observations each (Training sessions and Matchday).

Performance Metrics

Key performance metrics analyzed include:

- Number of high-intensity runs per minute (HIRpMIN)

- Distance covered during high-intensity runs per minute (DISTpMIN)

- Total distance covered per minute (TotalDISTpMIN)

- Number of sprints

- Sprint distance

Machine Learning Models

Five machine learning models were implemented for performance prediction:

1. Random Forest Regression

2. Decision Tree Regression

3. Gradient Boosting Regression

4. Lasso Regression

5. Artificial Neural Network (Deep Learning)

The machine learning models were trained on 70% of the dataset, with the remaining 30% reserved as a hold-out test set for performance evaluation. Each model was assessed using two standard regression metrics: Mean Squared Error (MSE), which quantifies the average squared difference between actual and predicted values, and the coefficient of determination (R^2), which measures the proportion of variance explained by the model.

Analysis

Comparison of Average Match Demands for Different Playing Positions

Most physical performance measures revolve around distances and high-speed runs, and this running behavior can be profiled at different intensities. For instance, in this study, high-intensity runs are movements between

19.8 kph and 25 kph, while movements over 25.1 kph are considered sprints visible on their heatmaps. A few performance metrics will be used for this research including the number of Sprint and High-intensity events as well as the Distance for High-intensity runs (HIR) and Sprints.

Therefore, to utilize as much of the data as possible in this analysis, the number of high-intensity runs per minute of play (HIRpMIN) and the distance covered during high-intensity runs per minute of play (DISTpMIN) have been normalized by the duration of activity for training or match session.

Table 1. Player Metrics Description

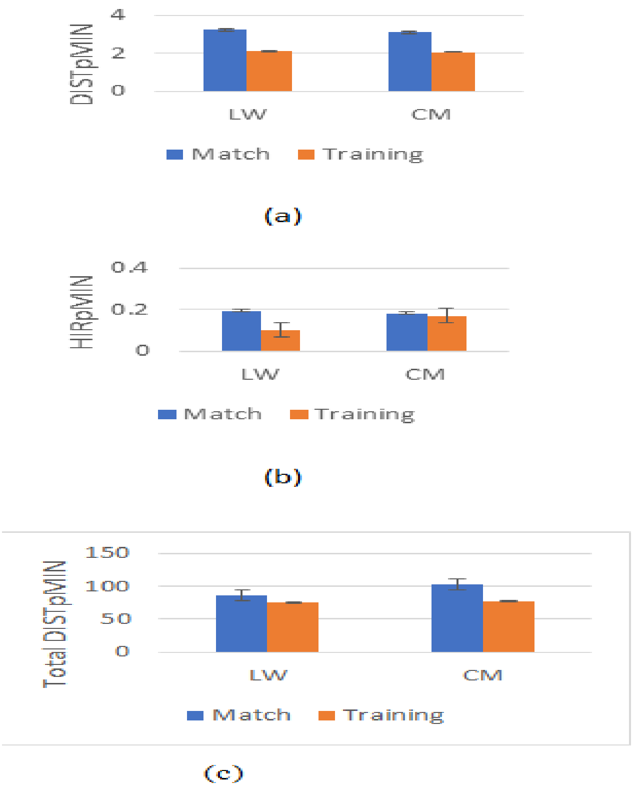
	SEASON 2022/2023	
Playing Position	**Metric**	**Match Average**
Left Winger (LW)	HIRpMIN	0.20
	DISTpMIN	3.25
	TotalDISTpMIN	86.56
Midfielder (CM)	HIRpMIN	0.18
	DISTpMIN	2.95
	TotalDISTpMIN	102.93

The seasonal variations can be observed for a given position as illustrated in the tables and diagrams. For instance, the average number of HIRs per minute in competitive matches for the Left winger (LW) at 0.20 in the 2022–2023 match season varies from that of the Central Midfielder (CM) at 0.18 in the same season. This process can be continuous for subsequent seasons to compare the physical demands of the seasons as well as oversee the improvement of players in terms of their physical contributions.

An indication from different observations shows the distance travelled during HIRs. For instance, across the season considered, the LW covered more distance at high Intensity or speeds and has more HIRs than the CM. However, the Total distance per minute (TotalDISTpMIN) variable is higher for the Midfielder as compared to the Left-wing position This shows that while the CM has the least number of high-intensity runs compared to the LW, on average the CM tends to cover more distance than the LW.

Physical Demands for Training and Match Sessions

The HIRpMIN variables show a significant difference between match and training session demands for all considered positions of play (LW: 1.195 in matches vs. 0.101 in training, and CM: 1.18 vs. 0.171 in training). Similarly, the distances covered during high intensity runs illustrate a significant difference between the match and training demands. The HIRpMIN variables show a significant difference between match and training sessions.



(a)

(b)

(c)

Fig 1. Comparison of average physical activity demand for match days and training sessions for different playing positions (a) DIST21pMIN; (b) HSR21pMIN; (c) Total distpmin

In Figure 3c, it is illustrated that the CM tends to cover the most distance (Total DISTpMIN) when compared to the LW position. The role of the central midfielder entails most contributions to the defensive and attacking phases of play. However, the physical demands which include the high-intensity distances (DISTpMIN) are quite similar, but the high-intensity runs (HIRpMIN) of both playing positions are a bit different with the frequency of the LW position being higher than the CM in match situations (Figure 3).

As explained, the nature of the CM position requires the player to significantly cover more distance in comparison with other positions [3] and agrees with the results of this study as illustrated in Figure 3 and Table 2. The results of this study show significant differences in physical demands in matches across the varying positions played. The difference in training sessions is not particularly significant because of the varying training drills and conditions.

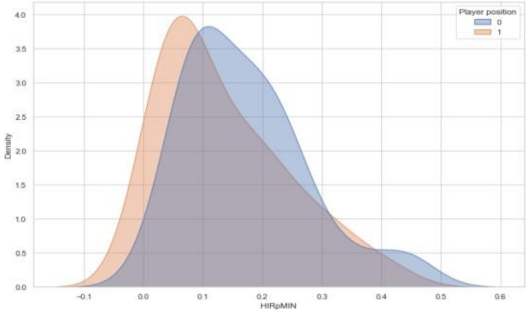


Fig 2. Comparison of the high-intensity runs by Player position across the season (Training and Match sessions), 0=LW, 1=CM

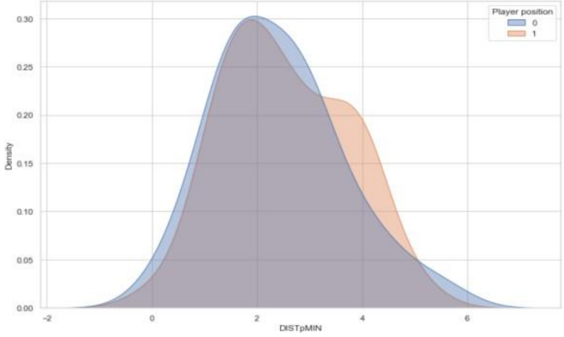


Fig 3. Comparison of the High-intensity distances by Player position across the season (Training and Match sessions), 0=LW, 1=CM

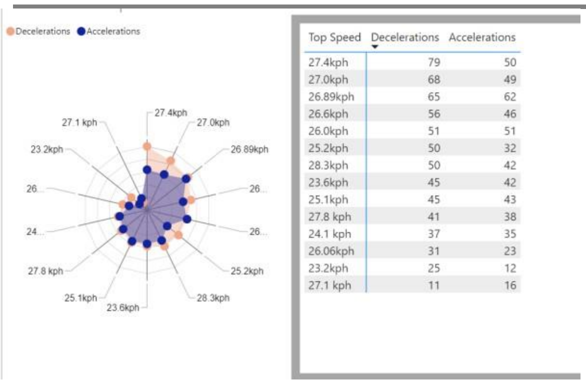
DISCUSSION

Informing Player and Team Analysis

Through analysis, it is deduced that from the player sample of this research, the midfielder had more physical output in terms of Distance covered throughout the season. A match average of 7100m (7.1km) was recorded in the first part of the season but improved to over 7300m (7.3km) in the second part of the season while the training session decreased from 4400m (4.4km) to 3700m (3.7km). Although various factors such as training types might have resulted in this, the match outcomes show a clear improvement, hence informing the player and team on player performance and helping to drive his development.

Detecting Injuries

By examining data from video recordings, sensors, and other sources, analytics, and artificial intelligence can be utilized to find football players who are at risk of injuries. A player's performance patterns, such as variations in speed, acceleration, and deceleration, which may be signs of an imminent injury, can be identified by Artificial intelligence algorithms useful to aid in predicting the likelihood of an injury and preventing burnout.



Top Speed	Decelerations	Accelerations
27.4kph	79	50
27.0kph	68	49
26.89kph	65	62
26.6kph	56	46
26.0kph	51	51
25.2kph	50	32
28.3kph	50	42
23.6kph	45	42
25.1kph	45	43
27.8 kph	41	38
24.1 kph	37	35
26.06kph	31	23
23.2kph	25	12
27.1 kph	11	16

Fig 4. Using analytics to detect injury-causing metrics.

According to Gronwald et al BJSM 2022, football players are prone to injuries that are sprint-related, accounting for 48% of all injuries sustained. These injuries commonly occur during linear acceleration, deceleration, and high-speed running. To prevent such injuries, analytical tools can be utilized to identify metrics that contribute to such occurrences.

Limitations and Future Work

While this study addresses certain needs, it is not without limitations. The comparison portrays a general seasonal comparison and is not specific to different types of competitions. Future research should consider longer-term data collection and additional physical factors to fully understand player performance and abilities.

RESULT AND EVALUATION

The machine learning models used in this regression model prediction include random forest (a supervised learning method, that uses the ensemble learning strategy for regression. By combining predictions from several machine learning algorithms, the ensemble learning technique produces predictions that are more accurate than those from a single model), a decision tree (builds a model in the shape of a tree to forecast future data and produces valuable continuous output by looking at an item's attributes), ridge regression model (a model-tuning technique used when analyzing data with multicollinearity. This method carries out L2 normalization. When the issue of multicollinearity, unbiased least squares, and significant variances occurs, the projected values diverge significantly from the actual values), gradient boosting, lasso regression model (a regularization method perfectly suited when a model shows a high level of multicollinearity or when you want to automate some processes in the model selection process), deep learning model, and artificial neural network.

From these models used for prediction, the Artificial Neural Network and Support Vector Regression Model have the highest accuracy and are best suited for this. While the Support Vector Regression has the lowest MSE. MSE (Mean Squared Error) measures the average squared difference between the predicted and actual values of the target variable. The Artificial Neural Network (ANN) used in this study consisted of two hidden layers with 64 neurons each, employing ReLU activation functions. The model was compiled using the Adam optimizer and trained for 100 epochs with a batch size of 32. The models were evaluated using the match day and training metrics producing a t-statistic: -8.623 and p-value: 1.08975e-13, since $p\text{-value} < 0.05$, the difference in Distance (km) between Training and Match Day is statistically significant. This means the mean distance in training is likely lower than Match Days

Table 2. Model Performance

Model	**Mean Squared Error**	**Average R²**
ANN	0.130	0.4636
Lasso Regression	0.149	0.0614

Gradient Boosting (GB)	0.1249	0.3255
Support Vector Regression (SVR)	0.0823	0.3810
Decision Tree (DT)	0.0968	0.3331
Random Forest (RF)	0.1020	0.4895

Table 3. Feature definition

Time Played	Duration or amount of time for training and match sessions for players
Distance (km)	The sum of the area covered during a game
Metre per Minute	The speed at which a player covers a distance of one meter per minute.
Top Speed	Maximum speed by a player during a session
HIRpMIN (High-Intensity Run per Minute)	Number of High-Intensity Runs / Time played
High Intensity Run(m)	Speed by a player over 19.8km
Player Position	Position of the player while on the pitch (Wing / Midfield)
Key	Training or Match Session; 0 and 1 respectively
Number of Sprints	Number of runs made at over 25.1km
Sprint Distance	Distance covered at a speed over 25.1km

CONCLUSION

This study demonstrates the potential of data analytics, machine learning, and deep learning in grassroots football performance analysis. By comparing high-intensity runs and distances during training and matches, we established significant differences in physical demands across playing positions. The implementation of machine learning algorithms for performance prediction shows promise in informing player development strategies.

The results highlight the importance of position-specific training and the need for tailored approaches to match preparation. Furthermore, the use of analytics for injury risk detection opens up new possibilities for player health management in grassroots football.

Future research should focus on long-term data collection, incorporation of additional performance factors, and exploration of other aspects of performance analysis, including physiological and tactical analyses.

ACKNOWLEDGMENT

First and foremost, I would like to express my sincere gratitude to my supervisor and co-supervisor, Alsmadi Hiba and Nawaz Mansha for their invaluable advice, guidance, and continuous support. I am also grateful to FC Pudsey who shared the pitch during this research, for their commitment and contribution during the 2022/2023 football season. Finally, I would like to express my gratitude to my parents, and friends for their tremendous understanding and support

REFERENCES

1. P.S. Bradley, W. Sheldon, B. Wooster, P. Olsen, P. Boanas, and P. Krustrup, "High intensity running in English FA Premier League soccer matches," Journal of Sports Sciences, vol. 27, no. 2, pp. 159-168, 2009.
2. C. Barnes, D.T. Archer, B. Hogg, M. Bush, and P.S. Bradley, "The evolution of physical and technical performance parameters in the English Premier League," International Journal of Sports Medicine, vol. 35, no. 13, pp. 1095-1100, 2014.
3. M. Bush, C. Barnes, D.T. Archer, B. Hogg, and P.S. Bradley, "Evolution of match performance parameters for various playing positions in the English Premier League," Human Movement Science, vol. 39, pp. 1-11, 2015.
4. Hewitt, K. Norton, and K. Lyons, "Movement profiles of elite women soccer players during international matches and the effect of opposition's team ranking," Journal of Sports Sciences, vol. 32, no. 19, pp. 1874-1880, 2014.
5. R.A. Zuber, J.M. Gandar, and B.D. Bowers, "Beating the spread: Testing the efficiency of the gambling market for National Football League games," Journal of Political Economy, vol. 93, no. 4, pp. 800-806, 2015.
6. N. Zaveri, S. Tiwari, P. Shinde, U. Shah, and L.K. Teli, "Prediction of football match score and decision-making process," International Journal of Recent Innovations in Trends in Computer and Communication, vol. 6, no. 2, pp. 162-165, 2018.
7. S.B. Maind and Ms. Sonali B., "Research paper on the basics of Artificial Neural Networks," International Journal of Recent Innovations in Trends in Computer and Communication, vol. 2, no. 1, pp. 96-100, 2014.
8. H. Liu, M.A. Gómez, C. Lago-Peñas, and J. Sampaio, "Match performance profiles of goalkeepers of elite football teams," International Journal of Sports Science and Coaching, vol. 10, no. 4, pp. 669-682, 2015.
9. D. Weimar and P. Wicker, "Moneyball revisited: Effort and team performance in professional soccer," Journal of Sports Economics, vol. 18, no. 2, pp. 140-161, 2017.
10. C. Petersen, D. Pyne, and B. Dawson, "Technology in sports performance analysis: An update," Sports Medicine, vol. 48, no. 7, pp. 1543-1558, 2018.
11. Petersen, C., Pyne, D., & Dawson, B. (2018). Technology in sports performance analysis: An update. Sports Medicine, 48(7), 1543-1558.
12. Liu, H., Gómez, M.A., Lago-Peñas, C., & Sampaio, J. (2019). Data analysis techniques in sports performance: A systematic review. Sports Medicine, 49(4), 453-484.
13. Rabitti, E., Boccia, G., Zangla, D., & Annino, G. (2020). Modeling techniques in sports performance analysis: An overview. Applied Sciences, 10(3), 850.