ORIGINAL ARTICLE

Machine learning analysis for predicting performance in female volleyball players in India: Implications for talent identification and player development strategies

Swamynathan Sanjaykumar. *Department of Physical Education and Sports Sciences. Faculty of Science and Humanities. SRM Institute of Science and Technology. Kattankulathur, India.*

Subhashree Natarajan. *School of Business. RV University. Karnataka, India.*

Ponnusamy Yoga Lakshmi ✉. *Department of Computer Science. Faculty of Science and Humanities. SRM Institute of Science and Technology. Kattankulathur, India.*

Yuliya Kalmykova. *Department of Physical Therapy. Kharkiv State Academy of Physical Culture. Kharkiv, Ukraine.*

Joseph Lobo. *College of Sports Exercise and Recreation. Bulacan State University. Bulacan, Philippines.*

Ratko Pavlović. *Faculty of Physical Education and Sport. University of East Sarajevo. Republic Srpska, Bosnia and Herzegovina.*

Edi Setiawan. *Faculty of Teacher Training and Education. Department of Physical Education, Health and Recreation. University Suryakancana. Cianjur, Indonesia.*

ABSTRACT

Talent identification and player development are crucial aspects of sports management, particularly in volleyball, where understanding players' performance predictors is essential. The primary objective is to investigate the relationships between players' demographic and physical attributes and their on-court performance, providing valuable insights for talent identification and player development strategies. The dataset comprises demographic and physical attributes alongside performance metrics of college-level female volleyball players in India. Data were meticulously collected from various institutions participating in volleyball tournaments across India. Three machine learning algorithms—linear regression, random forest regression, and XGBoost regression—were trained using the pre-processed dataset. Standard regression evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2) score were used to assess model performance. Random forest regression emerged as the top-performing ML technique, achieving a prediction accuracy of 94.18%, followed by XGBoost regression with 92.76%. Height, muscle mass, and bone mass exhibited strong positive correlations with performance prediction, emphasizing their significance. This study highlights ML techniques' potential, particularly random forest regression, in improving talent identification and performance prediction in college-level female volleyball players in India.

Keywords: Performance analysis, Volleyball, Performance prediction, Machine learning, Physical attributes, Talent identification.

Cite this article as:
Sanjaykumar, S., Lakshmi, P. Y., Natarajan, S., Kalmykova, Y., Lobo, J., Pavlović, R., & Setiawan, E. (2025). Machine learning analysis for predicting performance in female volleyball players in India: Implications for talent identification and player development strategies. *Journal of Human Sport and Exercise, 20*(1), 207-215. https://doi.org/10.55860/cn2vdj44

✉ **Corresponding author.** *Department of Computer Science, Faculty of Science and Humanities, SRM Institute of Science and Technology, Kattankulathur, India.*
E-mail: yogalakp@srmist.edu.in

Submitted for publication July 31, 2024.
Accepted for publication September 12, 2024.
Published October 30, 2024.
Journal of Human Sport and Exercise. ISSN 1988-5202.

doi: https://doi.org/10.55860/cn2vdj44

INTRODUCTION

Volleyball has emerged as a sport of significant interest and participation worldwide, including in India, where it is particularly popular at the college level. Globally, volleyball is not only a recreational activity but also a competitive sport that fosters community and athletic excellence (Koley, 2011; Alvares et al., 2023). College-level tournaments serve as crucial breeding grounds for budding talent, providing young athletes with platforms to hone their skills and showcase their potential (Tiaprapong & Tiaprapong, 2022). Identifying and nurturing promising players is essential for the sustained growth and success of volleyball. However, traditional talent identification methods often rely on subjective evaluations, which can overlook hidden talents or misjudge player capabilities, leading to missed opportunities for player development and hindering the overall progress of the sport (Śliwa et al., 2021; Römer et al., 2023). The significance of effective talent identification extends beyond India and is a global concern in the sports community. Various authors have highlighted the limitations of conventional evaluation techniques across different sports, emphasizing the need for more objective, data-driven approaches to better assess and predict athletic performance. Inaccurate assessments can result in the loss of potential elite athletes and impede the development of the sport on a broader scale (Kapadia et al., 2022; Sumathi et al., 2023).

Machine learning (ML) techniques offer a promising solution by providing a data-driven approach to talent evaluation and performance prediction. By analysing historical data, ML models can make objective assessments of player capabilities, potentially revolutionizing talent identification processes. Despite the widespread use of ML in sports such as football, basketball, and cricket for talent identification and performance prediction, its application in volleyball remains relatively underexplored, especially at the college level (Bunker & Thabtah, 2019; Passi & Pandey, 2018). This study aims to address this gap by investigating the effectiveness of ML techniques in predicting the future performance of college-level female volleyball players globally, with a focus on India as a case study (Van Eetvelde et al., 2021; Baboota & Kaur, 2019). This research will explore the use of three main ML techniques: linear regression, random forest regression, and XGBoost regression. These techniques are chosen for their versatility, interpretability, and ability to manage both numerical and categorical data. By analysing past performance metrics, physical attributes, and other relevant factors, these ML models can provide valuable insights into the future potential of players (de Leeuw et al., 2022; Ju & Zhang, 2023). The primary objective of this research is to obtain new knowledge regarding the application of ML in volleyball talent identification and performance prediction (Śliwa et al., 2021; Abebe et al., 2020).

MATERIALS AND METHODS

Participants

The dataset for this study encompasses demographic and physical attributes of 174 college-level female volleyball players aged between 18 and 25 years in India, along with performance metrics recorded during matches or training sessions. The data were collected from various colleges and institutions participating in volleyball tournaments across different regions of India. Attributes such as age, height, weight, body fat percentage, fat mass, muscle mass, bone mass, BMI, and performance indicators were meticulously recorded for each player.

Procedures

The dataset underwent preprocessing to ensure data quality and consistency, including handling missing values, outliers, and data normalization for feature scaling (Figure 1). Categorical variables were encoded as numerical values if applicable. Three machine learning algorithms—linear regression, random forest

regression, and XGBoost regression—were implemented to predict the performance of college-level female volleyball players based on their demographic and physical attributes (Baboota & Kaur, 2019; Hu et al., 2022). Each algorithm was trained using the pre-processed dataset, with performance metrics serving as the target variable. Model evaluation was conducted using cross-validation techniques to assess predictive accuracy and generalization ability (Miguel-Ortega et al., 2023; Hudnurkar & Rayavarapu, 2022).

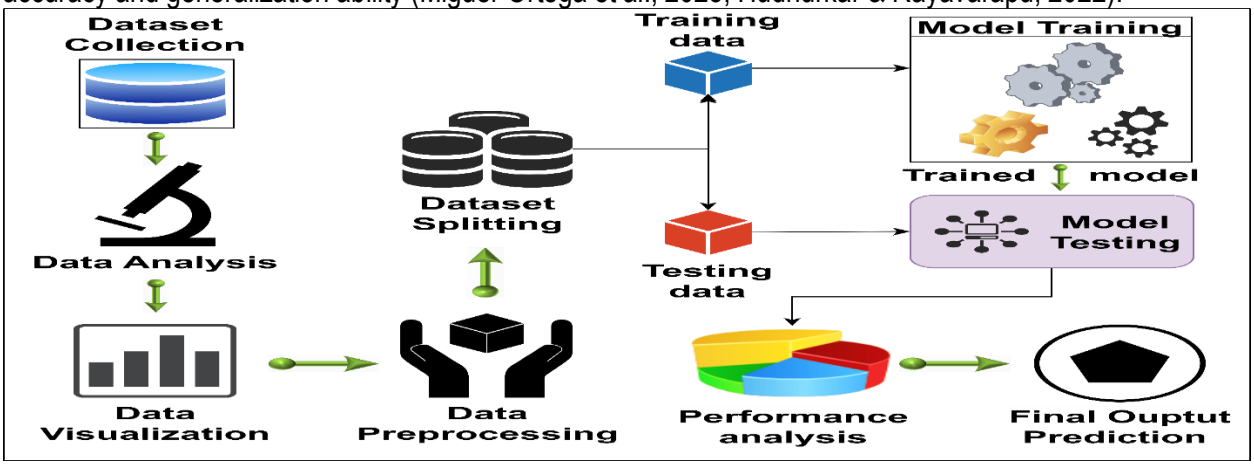

Figure 1. Diagram that represents the data processing architecture.

Statistical analysis

The performance of each model was evaluated using standard regression evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared. These metrics provide insights into the models' ability to accurately predict player performance based on the input features. Statistical tests, such as Pearson correlation analysis, were performed to examine the relationships between demographic and physical attributes and player performance (Lakshmi et al., 2024; Sanjaykumar et al., 2024). This analysis helps in identifying significant predictors and understanding the underlying patterns in the data.

RESULTS

Table 1 presents a sample dataset of college-level female volleyball players in India, showcasing their demographic and physical attributes, including age, height, weight, body fat percentage, fat mass, muscle mass, bone mass, BMI, and performance prediction (PP). These attributes provide insights into the characteristics of the participants in the study.

Table 1. Sample dataset of college-level female volleyball players in India.

Number of participants (n = 174)	Age (years)	Height (cm)	Weight (kg)	Fat (%)	Fat mass (kg)	Muscle mass (kg)	Bone mass (kg)	BMI (kg/m^2)	PP (%)
32	19	170	85.2	28.8	30.3	65.6	3.3	28.9	78
54	20	178	83.8	28.4	36.3	64.9	3.9	26.1	81
37	21	185	85.2	22.3	19	62.8	3.4	23.3	85
34	22	172	88	30.3	39	67.3	3.7	31.9	73
17	23	185	80.8	21.6	20.3	59	3.5	21.5	91

Note. cm = centimetre, kg = kilograms, % = percentage, m = meter.

Table 2. Performance comparison of machine learning algorithms for predicting volleyball player performance.

Algorithm	Mean Squared Error (MSE)	R-squared Score	Root Mean Squared Error (RMSE)
Linear Regression	30.3207	0.7531	5.5064
Random Forest	7.1505	0.9418	2.6740
XGBoost	8.8876	0.9276	2.9812

Performance prediction based on Linear Regression model

The linear regression model offers a reasonable approach to predicting the performance of college-level female volleyball players in India, with an achieved prediction accuracy of 75.31%. While the model provides a foundation for understanding the relationship between input features and player performance. With a moderate R-squared score of 0.7531, the model explains approximately 75.31% of the variability in player performance. However, the relatively high Mean Squared Error (MSE) of 30.3207 and Root Mean Squared Error (RMSE) of 5.5064 suggest notable prediction errors (Table 2). Despite its limitations, linear regression can still offer valuable insights into player performance, serving as a baseline for more sophisticated modelling techniques. Figure 2 illustrates the graphical representation of the linear regression model used to predict volleyball players' performance. The plot showcases the relationship between the players' demographic and physical attributes and their overall performance scores, providing insights into the predictive capability of the model.

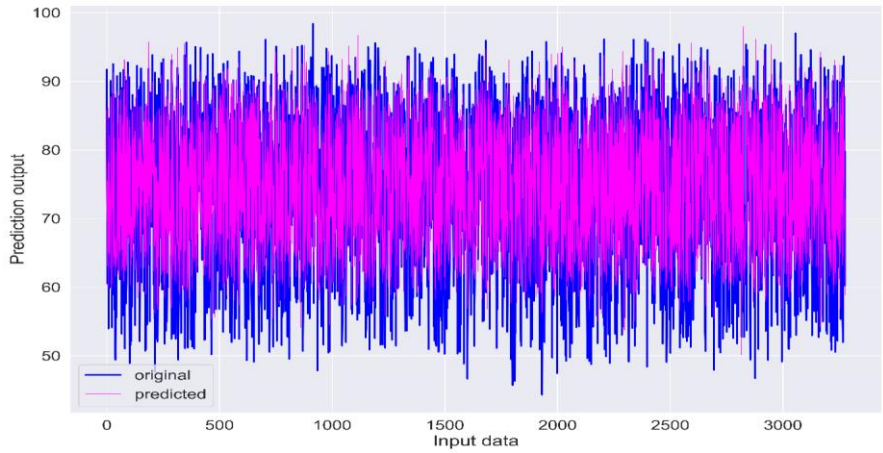

Figure 2. The linear regression model's graphical representation for predicting volleyball players' performance, the graph shows the model's predictions (pink) against input data (blue).

Performance prediction based on Random Forest Regression model

Random forest regression emerges as a highly effective approach for predicting the performance of college-level female volleyball players in India, boasting an impressive prediction accuracy of 94.18%. The model achieves a significantly lower Mean Squared Error (MSE) of 7.1505 and a high R-squared score of 0.9418, indicating its capability to explain approximately 94.18% of the variability in player performance. With a Root Mean Squared Error (RMSE) of 2.6740, the model's prediction errors are notably reduced compared to linear regression (Table 2). The ensemble nature of random forest regression, combining multiple decision trees, allows it to capture complex relationships within the data effectively. This robust performance makes random forest regression a valuable tool for talent identification and player development in the realm of volleyball. Figure 3 illustrates the graphical representation of the Random Forest Regression model used to predict

volleyball players' performance. The plot showcases the relationship between the players' demographic and physical attributes and their overall performance scores, providing insights into the predictive capability of the model.

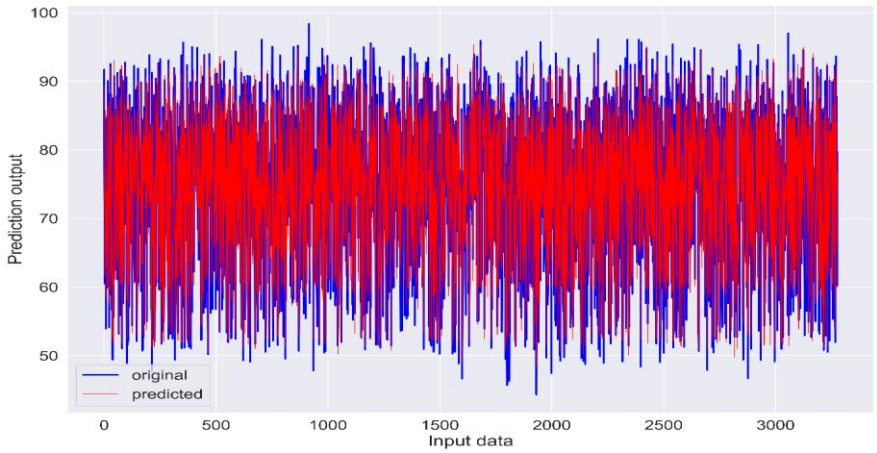

Figure 3. The random forest model's graphical representation for predicting volleyball players' performance, the graph shows the model's predictions (red) against input data (blue).

Performance prediction based on XGBoost Regression model

XGBoost regression emerges as a formidable contender in predicting the performance of college-level female volleyball players in India, achieving a prediction accuracy of 92.76%. With a relatively low Mean Squared Error (MSE) of 8.8876 and a high R-squared score of 0.9276, the model explains approximately 92.76% of the variability in player performance. The Root Mean Squared Error (RMSE) of 2.9812 indicates a reduced average prediction error compared to linear regression (Table 2). XGBoost's gradient boosting framework enables sequential learning, allowing the model to capture intricate patterns in the data and make accurate predictions. Despite slight differences in performance compared to random forest regression, XGBoost regression remains a powerful tool for talent identification and player development initiatives in volleyball. Figure 4 the graphical representation of the XGBoost Regression model for predicting volleyball players' performance. It highlights the relationship between players' attributes and performance scores, offering insights into the model's predictive capability.

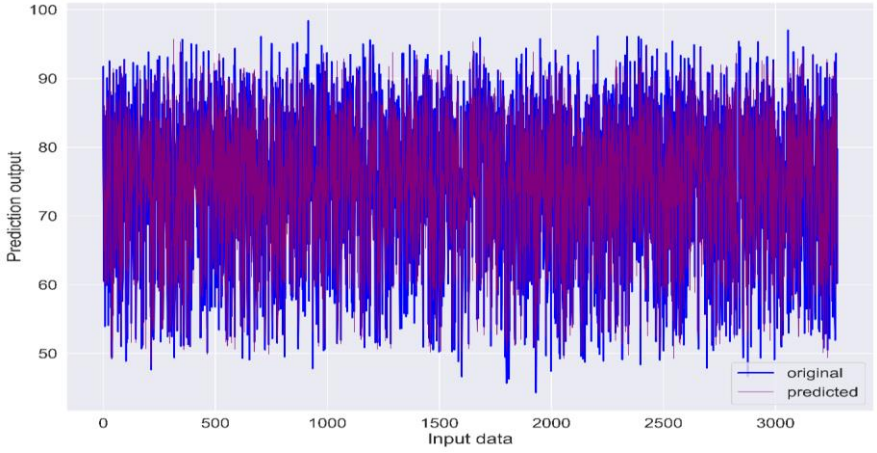

Figure 4. The XGBoost regression model's graphical representation for predicting volleyball players' performance, the graph shows the model's predictions (maroon) against input data (blue).

Table 3. Pearson correlation matrix for attributes and performance prediction.

Attribute	Age	Height	Weight	Fat	Fat mass	Muscle mass	Bone Mass	BMI	Performance Prediction
Age (year)	1	0.1151	0.4054	0.4759	0.4184	0.2644	0.2697	0.3881	0.2771
Height (cm)	0.1151	1	0.1245	-0.1091	-0.0381	0.5588	0.5332	-0.2217	0.8790
Weight (kg)	0.4054	0.1245	1	0.9358	0.9784	0.8052	0.8178	0.9383	0.3379
Fat (%)	0.4759	-0.1091	0.9358	1	0.9768	0.5778	0.5969	0.9627	0.1711
Fat mass (kg)	0.4184	-0.0381	0.9784	0.976	1	0.6650	0.6835	0.9737	0.1982
Muscle mass (kg)	0.2644	0.5588	0.8052	0.5778	0.6650	1	0.9920	0.5960	0.6525
Bone Mass (kg)	0.2697	0.5332	0.8178	0.5969	0.6835	0.9920	1	0.6166	0.6222
BMI (kg/m^2)	0.3881	-0.2217	0.9383	0.9627	0.9737	0.5960	0.6166	1	0.0416
Performance Prediction (%)	0.2771	0.8790	0.3379	0.1711	0.1982	0.6525	0.6222	0.0416	1

Note. cm = centimetres, kg = kilograms, % = percentage, BMI = Body Mass Index, kg/m^2 = kilograms/meters square.

The results indicate that several physical attributes are significantly correlated with performance prediction. Height shows a strong positive correlation (0.8790, $p < .001$) with performance prediction, suggesting that taller players tend to perform better. Muscle mass (0.6525, $p < .001$) and bone mass (0.6222, $p < .001$) also exhibit strong positive correlations with performance prediction, indicating their importance in evaluating player performance. Attributes such as age and BMI show weaker correlations with performance prediction, with BMI exhibiting a non-significant correlation (0.0416, $p = .5857$). This suggests that while BMI might not significantly impact performance prediction, other factors like height, muscle mass, and bone mass are crucial (Table 3).

DISCUSSION

The present study aimed to utilize machine learning (ML) algorithms to predict the performance of college-level female volleyball players in India by considering a range of demographic and physical attributes. The research question posited whether ML models could enhance the accuracy and objectivity of talent evaluations compared to traditional methods (Abebe et al., 2020; Borowiec et al., 2023).

The results indicate that all three ML models—linear regression, random forest regression, and XGBoost regression—demonstrate varying degrees of predictive efficacy. Specifically, random forest regression emerged as the top performer, achieving an impressive prediction accuracy of 94.18%, as evidenced by its high R-squared score of 0.9418 and low Mean Squared Error (MSE) of 7.1505. This model's ensemble nature, which combines multiple decision trees, effectively captures complex relationships within the data, making it particularly suited for predicting volleyball player performance (Bai & Bai, 2021). The hypothesis that ML models can significantly enhance the accuracy of talent evaluations is confirmed by these results. The random forest regression model, in particular, significantly outperforms linear regression, which achieved a moderate prediction accuracy of 75.31%, with a higher MSE of 30.3207 and an R-squared score of 0.7531. XGBoost regression also performed well, with a prediction accuracy of 92.76%, demonstrating its capacity to capture intricate data patterns, although slightly less effectively than random forest regression (Alvares et al., 2023). The Pearson correlation analysis provided valuable insights into the relationships between physical attributes and performance prediction. Height demonstrated a strong positive correlation (0.8790, $p < .001$) with performance prediction, indicating that taller players tend to perform better. This finding is consistent with the established understanding that height can be a critical factor in volleyball, affecting a player's ability to spike, block, and reach. Muscle mass (0.6525, $p < .001$) and bone mass (0.6222, $p < .001$) also showed strong positive correlations with performance prediction. These attributes are indicative of

physical strength and endurance, which are essential for powerful plays and sustained performance throughout a match. The strong correlations highlight the importance of these physical characteristics in evaluating volleyball players. On the other hand, attributes such as age and BMI exhibited weaker correlations with performance prediction. Specifically, BMI had a non-significant correlation (0.0416, p = .5857), suggesting it might not be a reliable predictor of performance. This may be due to BMI's inability to differentiate between muscle and fat mass, which are crucial in athletic performance contexts (Römer et al., 2023; Miguel-Ortega et al., 2023). However, it is crucial to acknowledge that these models focused solely on these attributes, potentially overlooking other critical factors such as skill level, experience, and psychological aspects, which could further enhance prediction accuracy. The study successfully demonstrates the potential of ML techniques in improving talent identification and performance prediction in volleyball. By providing a data-driven approach, ML models, particularly random forest regression, can offer valuable insights for coaches, scouts, and policymakers in the volleyball community. Future research should aim to incorporate a broader range of variables, including skill and psychological factors, and explore the applicability of these techniques across different demographics and levels of competition to generalize these findings further (Sanjaykumar et al., 2024; López-Serrano et al.,2024; Šuštaršič et al., 2022).

CONCLUSIONS

This study showcases the potential of machine learning techniques, particularly random forest regression, in enhancing talent identification and performance prediction in college-level female volleyball players in India. The results demonstrate that physical attributes such as height, muscle mass, and bone mass play significant roles in predicting player performance. While traditional methods rely on subjective evaluations, the application of data-driven approaches offers more accurate and objective insights. Future research should focus on incorporating additional variables, including skill level and psychological factors, to further refine predictive models and generalize findings across diverse demographics and competition levels, ultimately advancing talent development in volleyball.

AUTHOR CONTRIBUTIONS

Swamynathan Sanjaykumar: conceptualization, methodology, investigation, resources, writing-review & editing, supervision, writing-original draft and final approval the manuscript. Subhashree Natarajan: resources, formal analysis, data curation and collection, supervision and final approval the manuscript. Ponnusamy Yoga Lakshmi: methodology, formal analysis, data curation and collection, writing-review & editing and final approval the manuscript. Yuliya Kalmykova: writing-review & editing, validation, supervision and final approval the manuscript. Joseph Lobo: formal analysis, supervision and final approval the manuscript. Ratko Pavlović: writing-review & editing, formal analysis, and final approval the manuscript. Edi Setiawan: Validation, formal analysis, and final approval the manuscript.

SUPPORTING AGENCIES

No funding agencies were reported by the authors.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

ACKNOWLEDGMENTS

We would like to express our gratitude to all the participants who contributed to this study by providing their data and insights. We also extend our appreciation to the institutions and organizations that supported and facilitated the data collection process.

REFERENCES

Abebe, M., Shin, Y., Noh, Y., Lee, S., & Lee, I. (2020). Machine learning approaches for ship speed prediction towards energy efficient shipping. Applied Sciences (Switzerland), 10(7), 2325. https://doi.org/10.3390/app10072325

Alvares, L. A. M., Ferreira, R. E. S., Nakamoto, F. P., et al. (2023). Physical fitness, hormonal profile, nutritional and psychological aspects assessment of transgender women volleyball players submitted to physical tests: protocol paper of a prospective cohort. BMJ Open Sport & Exercise Medicine, 9, e001641. https://doi.org/10.1136/bmjsem-2023-001641

Baboota, R., & Kaur, H. (2019). Predictive analysis and modeling football results using a machine learning approach for the English Premier League. International Journal of Forecasting, 35(2), 741-755. https://doi.org/10.1016/j.ijforecast.2018.01.003

Bai, Z., & Bai, X. (2021). Sports Big Data: Management, Analysis, Applications, and Challenges. Complexity. https://doi.org/10.1155/2021/6676297

Borowiec, J., Banio-Krajnik, A., Malchrowicz-Mośko, E., et al. (2023). Eating disorder risk in adolescent and adult female athletes: the role of body satisfaction, sport type, BMI, level of competition, and training background. BMC Sports Science, Medicine and Rehabilitation, 15, 91. https://doi.org/10.1186/s13102-023-00683-7

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. Applied Computing and Informatics, 15(1), 27-33. https://doi.org/10.1016/j.aci.2017.09.005

de Leeuw, A. W., van Baar, R., Knobbe, A., & van der Zwaard, S. (2022). Modeling Match Performance in Elite Volleyball Players: Importance of Jump Load and Strength Training Characteristics. Sensors, 22(20), 7996. https://doi.org/10.3390/s22207996

Hu, L., Zhao, K., & Jiang, W. (2022). Biomechanical analysis of volleyball players' spike swing based on deep learning. Computational Intelligence and Neuroscience. https://doi.org/10.1155/2022/4797273

Hudnurkar, S., & Rayavarapu, N. (2022). Binary classification of rainfall time-series using machine learning algorithms. International Journal of Electrical and Computer Engineering, 12(2), 1945-1954. https://doi.org/10.11591/ijece.v12i2.pp1945-1954

Ju, H., & Zhang, H. (2023). Application of Multiple Linear Regression Model in the Sustainable Development of National Traditional Sports. Applied Mathematics and Nonlinear Sciences, 8(2), 3033-3042. https://doi.org/10.2478/amns.2023.2.00019

Kapadia, K., Abdel-Jaber, H., Thabtah, F., & Hadi, W. (2022). Sport analytics for cricket game results using machine learning: An experimental study. Applied Computing and Informatics, 18(3-4), 256-266. https://doi.org/10.1016/j.aci.2019.11.006

Koley, S. (2011). Correlations of Handgrip Strength with Selected Hand-Arm-Anthropometric Variables in Indian Inter-university Female Volleyball Players. Asian Journal of Sports Medicine, 2(4), 347-38. https://doi.org/10.5812/asjsm.34738

Lakshmi, P. Y., Sanjaykumar, S., Dharuman, M., & Elangovan, A. (2024). Using Support Vector Regression Kernel Models for Cricket Performance Prediction in the Womens Premier League 2024. Physical Education Theory and Methodology, 24(1), 72-78. https://doi.org/10.17309/tmfv.2024.1.09

López-Serrano, C., Hernández Gonzalez, C., Sánchez Morillas, P., López, E., & Molina Martín, J. J. (2024). Impact of early leadership on performance in volleyball sets. Journal of Human Sport and Exercise, 19(4), 992-1008. https://doi.org/10.55860/aybbzk53

Miguel-Ortega, Á., Calleja-González, J., & Mielgo-Ayuso, J. (2023). Comparison of Sports Performance and Kinanthropometric Profiles of Elite Female Basketball and Volleyball Players over the Course of a Competitive Season. Applied Sciences, 13(14), 8267. https://doi.org/10.3390/app13148267

Passi, K., & Pandey, N. (2018). Increased Prediction Accuracy in the Game of Cricket Using Machine Learning. International Journal of Data Mining & Knowledge Management Process, 8, 19-36. https://doi.org/10.5121/ijdkp.2018.8203

Römer, C., Zessin, E., Czupajllo, J., Fischer, T., Wolfarth, B., & Lerchbaumer, M. H. (2023). Effect of Physical Parameters and Training Load on Patellar Tendon Stiffness in Professional Athletes. Diagnostics, 13(15), 2541. https://doi.org/10.3390/diagnostics13152541

Sanjaykumar, S., Natarajan, S., Lakshmi, P. Y., & Boby, F. A. (2024). Predicting Team Success in the Indian Premier League Cricket 2024 Season Using Random Forest Analysis. Physical Education Theory and Methodology, 24(2), 304-309. https://doi.org/10.17309/tmfv.2024.2.16

Sanjaykumar, S., Udaichi, K., Rajendiran, G., Cretu, M., & Kozina, Z. (2024). Cricket performance predictions: a comparative analysis of machine learning models for predicting cricket player's performance in the One Day International (ODI) world cup 2023. Health, Sport, Rehabilitation, 10(1), 6-19. https://doi.org/10.58962/HSR.2024.10.1.6-19

Śliwa, M., Sadowski, J., & Buszta, M. (2021). Relative Age Effect and Talent Identification in Youth Volleyball Players from the Polish Volleyball Federation Sports School. Polish Journal of Sport and Tourism, 28(4), 21-25. https://doi.org/10.2478/pjst-2021-0022

Sumathi, M., Prabu, S., & Rajkamal, M. (2023). Cricket Players Performance Prediction and Evaluation Using Machine Learning Algorithms. In Proceedings of the 1st IEEE International Conference on Networking and Communications 2023, ICNWC 2023. https://doi.org/10.1109/ICNWC57852.2023.10127503

Šuštaršič, A., Videmšek, M., Karpljuk, D., Miloloža, I., & Meško, M. (2022). Big Data in Sports: A Bibliometric and Topic Study. Business Systems Research, 13(1), 19-34. https://doi.org/10.2478/bsrj-2022-0002

Tiaprapong, K., & Tiaprapong, K. (2022). The Relationship between Respiratory Muscle Strength and Physical Performance in College Volleyball Players. Sport Mont, 20(2), 41-45. https://doi.org/10.26773/smj.220607

Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. Journal of Experimental Orthopaedics, 8(1), 27. https://doi.org/10.1186/s40634-021-00346-x