Predictive Modelling of Player Performance in the Indian Super League Using Publicly Available Match Data: A Machine Learning Approach

Dr. Surekha S. Daptare[1], Dr. Vidya Dattatray Pathare[2], Dr. P.K. Lohote[3], Dr. Dnyaneshwar Pandurang Chimate[4], Dr. Anjushree Anthony Augustine[5], Anthony Augustine[6], Dr. Dinesh Eknath Ukirde[7], Dr. R. R. Chavan[8], Dr. S. Sujanesh K. Das[9], Dr. George Abraham[10]

[1]Director of Physical Education, Department of Sports and Physical Education, MGV's Mahilaratna Pushpatai Hiray Arts, Science and Commerce Mahila Mahavidyalaya, Malegaon Camp, Nashik, Maharashtra, India, Email: surekha25881@gmail.com
ORCID: https://orcid.org/0000-0001-9171-9472
[2]Director of Physical Education and Sports, Department of Physical Education and Sports, PDEA's Baburaoj Gholap College, Sangvi, Pune, Maharashtra, India.
Email: vidya.pathare30@gmail.com
[3]Director of Physical Education and Sports, Department of Physical Education and Sports, Mahatma Phule Mahavidyalaya, Pimpri, Pune, Maharashtra, India,
Email: pklohote@gmail.com
[4]Director of Physical Education and Sports, Department of Physical Education and Sports, PDEA's Prof. Ramakrishna More Arts, Commerce and Science College, Akurdi, Pune, Maharashtra, India, Email: rmcsports07@gmail.com
[5]Director of Physical Education and Sports, Department of Physical Education and Sports, MKSSS's Cummins College of Engineering for Women, Pune, Maharashtra, India.
Email: anjushree.augustine@cummiscollege.in
[6]Assistant Director of Physical Education and Sports, Department of Sports, Recreation & Wellness, Symbiosis International University, Pune, Maharashtra, India.
[7]Director of Physical Education and Sports, Department of Physical Education and Sports, MVP's Arts, Commerce and Science College, Tryambkeshwar, Nashik, Maharashtra, India.
Email: dineshukirde@gmail.com
[8]Director of Physical Education and Sports, Department of Physical Education and Sports, MVP's Arts, Commerce and Science College, Dindori, Nashik, Maharashtra, India.
chavanravir@gmail.com
[9]Associate Professor and Head, Department of Physical Education, Government Medical College, Calicut, Kerala, India. Email: drsujaneshphysical@gmail.com
[10]Professor, YMCA College of Physical Education, Chennai, Tamil Nadu, India.
Email: profgeorgeabraham@gmail.com ORCID: https://orcid.org/0000-0002-9763-555X

ABSTRACT

The global football leagues have widely adopted these techniques; Indian football is gradually catching up. In this paper, we present a machine learning-based approach specifically tailored for the Indian Super League (ISL) to predict individual player performance using publicly available match data. The collected and curated event-level data from over 300 ISL matches, focusing on features such as passes completed, tackles made, shots on target, and minutes played. From this data, we developed two predictive models: a regression model to estimate

expected goals (xG) for each player, and a classification model to predict the likelihood of a player being named 'Man of the Match'. These models were trained and tested using standard machine learning techniques including Random Forest and Logistic Regression, and achieved encouraging accuracy and consistency. The results highlight that even with limited but structured data; it is possible to uncover meaningful insights into player contributions. This work serves as a step toward bridging the gap between traditional sports analysis and modern data-driven methods in Indian football. This study approach is scalable, accessible, and adaptable for teams, coaches, and analysts aiming to adopt a more objective and data-informed strategy.

KEYWORDS: Indian super league, sports analytics, player performance, machine learning, expected goals (xG), predictive modelling, ISL data, man of the match.

INTRODUCTION

Football, or as it's fondly referred to as in lots of parts of India 'football' has long held the power to stir emotions, deliver groups together, and create heroes on the sphere. Whilst the lovely game has a wealthy international history of tactical innovation and participant improvement, India has handiest these days started out tapping into the full-size capability of records-driven football evaluation. The Indian remarkable League (ISL), released in 2014, has been at the leading edge of popularizing expert football in India, drawing lovers, media, and expertise from all corners of the country, but, in terms of reading participant overall performance with the precision visible in European or American leagues, there remains a sizable gap. In the worldwide football landscape, data analytics has converted from a behind-the-scenes guide tool to a relevant pillar of selection-making, influencing the whole thing from team selection to match approach and player transfers. Club in leagues just like the English most advantageous League, la Liga, and Bundesliga have adopted sophisticated metrics like anticipated goals (xG), possession value models, and player warmth maps to benefit a competitive facet. In competitive sports both capacities are very important, the first one is cardiovascular stamina (Tyagi et al., 2025) and the second one is muscle size and strength (Abraham, 2010). The football game skill is usually by the lower extremities and the size of the muscle (Sankaranarayanan et al., 2011) is also play a role to improve the skills. In assessment, Indian soccer has been slower to adopt those tools, partly due to the restrained availability of established suit records and a loss of reachable analytical frameworks tailor-made to the neighborhood context.

This study takes a step in that route. by way of focusing at the Indian top notch League and leveraging publicly available event-stage suit records, we purpose to show how even a modest dataset whilst curated and analyzed thoughtfully can reveal essential insights into player overall performance. We used statistics from over 300 ISL fits, taking pictures key on-discipline movements which include passes, tackles, shots, and playtime. Through this, we evolved center predictive models. The first estimates the expected dreams (xG) for man or woman gamers the use of regression techniques, supplying a measure of scoring efficiency past raw goal tallies. The second one is a category version that predicts whether a participant is likely to be named 'man of the healthy', a subjective but extensively accompanied accolade in football discussions.

This paper isn't pretty much models and metrics; it's approximately making analytics extra inclusive and grounded within the Indian soccer environment. We consider that democratizing get entry to such gear can empower coaches, analysts, or even fans to higher apprehend the

sport. Extra importantly, it shows that Indian football often underestimated on the worldwide stage has the potential to integrate superior techniques to enhance participant improvement and team strategy. The motivational level and metal abilities are also a vital role to improve the performance of game players (Justus & George, 2021; Thomas et al., 2024); this is not expected in football players (Abhinav & Abraham, 2022). In essence, this study tries to bridge the divide among and precision, between the raw energy of football and the quiet energy of numbers. As Indian football grows, so must its strategies of expertise overall performance. And this paper is a humble contribution to that evolving story.

Reviews

The intersection of sports and data science has gained immense traction over the past decade, with football leading the way in many innovations. Predictive modeling, player tracking, and performance evaluation systems have matured significantly in professional leagues across Europe and the Americas. One of the most studied metrics in football analytics is the "Expected Goals" (xG) model, first popularized in European football analysis. Studies such as those by Lucey et al. (2015) and Decroos et al. (2019) have introduced comprehensive frameworks to quantify scoring opportunities based on spatial and temporal features of on-pitch events. These models have been used not only to evaluate player efficiency but also to assess overall team strategy and tactical effectiveness. Beyond xG, machine learning has been applied to predict match outcomes, player injuries, team formations, and even crowd sentiment. Classifiers such as Random Forest, Support Vector Machines, and ensemble models have frequently been used due to their flexibility and interpretability in handling complex sporting data. Several works have also explored the prediction of accolades such as "Man of the Match", which is typically influenced by both quantifiable metrics and subjective assessments making it a challenging and nuanced task for machine learning algorithms.

When we narrow our lens to Indian football, the research landscape becomes significantly thinner. While the Indian Super League (ISL) has grown rapidly in viewership and investment since its inception in 2014, there has been a noticeable lack of academic or industry-grade analytical studies focused on its data. Most existing analysis has been descriptive in nature focused on league tables, team win rates, and individual top scorers. There are limited publicly documented attempts to apply predictive modeling or data science methodologies to ISL player performance data.

Table 1: All-Time Performance Summary of ISL Teams

Squad	MP	W	D	L	GF	GA	GD	Pts
Bengaluru	117	54	24	39	172	143	29	186
Chennaiyin	138	46	41	51	181	195	-14	179
ATK Mohan Bagan	92	45	22	25	135	100	35	157
Goa	156	71	41	44	285	204	81	254

Hyderabad	81	33	23	25	111	86	25	122
Jamshedpur	113	37	39	37	127	126	1	150
Kerala Blasters	144	41	46	57	161	200	-39	169
Mumbai City	142	67	32	43	243	169	74	233
North East United	144	34	43	67	140	219	-79	145
Odisha	95	26	23	46	129	167	-38	101
Pune City	69	23	19	27	83	91	-8	88
ATK	94	34	28	32	122	116	6	130
Delhi Dynamos	74	22	26	26	103	105	-2	92
FC Pune City	60	16	15	29	62	97	-35	63

The dataset we used, originally compiled by Suraj Suresh and hosted on Kaggle, are one of the few open-access, aggregated sources of ISL statistics spanning nearly a decade (2014–2023). It provides valuable match-level data such as passes, tackles, assists, shots, cards, and minutes played. Despite its simplicity compared to advance tracking datasets used in elite European leagues, it offers a solid foundation for developing predictive models tailored to Indian football. To our knowledge, few published studies have employed this dataset to build machine learning models, particularly ones that aim to predict individual player accolades or performance metrics like xG in an Indian context.

Thus, this research contributes to a relatively underexplored domain by demonstrating how standard machine learning techniques can be effectively applied to publicly available ISL data. In doing so, it not only offers methodological insights but also emphasizes the potential for grassroots-level analytics in Indian football, encouraging future studies in this growing space.

Dataset Description

The foundation of this research lies in a structured dataset sourced from publicly available Indian Super League (ISL) statistics, compiled by Suraj Suresh and hosted on Kaggle. The dataset, titled "Indian Super League Statistics (2014–2023)", brings together player-level and match-level information spanning across all ten seasons of the league, making it one of the most comprehensive open datasets available for Indian football.

The specific file used in this study All Time Combined.csv contains aggregated statistics for individual players across different seasons and matches. Each row in the dataset represents a

unique player-match instance, allowing us to analyze player performance over time and across teams. In total, the dataset includes data from over 300 ISL matches, covering thousands of player appearances. Key features in the dataset include:
Player Details: Name, club, nationality, and position (e.g., midfielder, forward, defender, goalkeeper). Match Statistics: Minutes played, goals scored, assists, shots taken, shots on target, passes completed, tackles made, yellow and red cards. Man of the Match: A binary indicator marking whether the player was awarded 'Man of the Match' for that particular game. Cumulative Metrics: For some players and matches, additional fields such as match date, season, and total appearances are also included
These features form the basis for building both regression and classification models in this research. To predict expected goals (xG), we primarily utilize offensive features such as shots, shots on target, and assists, while for the Man of the Match classification, a broader set of features both offensive and defensive are considered, including passes, tackles, cards, and playing time.
One of the challenges of working with this dataset is the absence of granular event data like player positions on the field, movement trajectories, or possession chains, which are typically available in advanced football analytics used in European leagues. However, despite these limitations, the dataset provides a reliable and rich source for modeling key performance indicators (KPIs) in the ISL context. It also reflects the level of data accessibility currently available in Indian football, reinforcing the importance of working with what is realistically obtainable. Before modeling, the dataset underwent a thorough cleaning and preprocessing phase to handle missing values, normalize formats, and engineer relevant features, which will be detailed in the following section.

MATERIALS AND METHODS

To predict player performance in the Indian Super League, our methodology is centered on two distinct yet complementary predictive modeling tasks: (1) estimating a player's expected goals (xG) using regression and (2) predicting the likelihood of a player being awarded 'Man of the Match' using classification. Each of these tasks involved careful data preparation, feature selection, model training, and evaluation using established machine learning practices. The raw dataset, although structured, required a systematic cleaning process to prepare it for model training. First, we filtered the data to retain only complete player-match entries, excluding rows with missing values in key performance fields such as shots, passes, or minutes played. Non-numerical attributes such as player names and club names were retained only for reference and later excluded from the modeling stage. Categorical features like player positions (e.g., defender, midfielder) were encoded using one-hot encoding. Numerical features such as goals, assists, passes, and tackles were standardized to ensure uniform scaling across the model. To mitigate the effect of extreme outliers, we applied a log transformation to features like total shots and minutes played when necessary.

Feature Engineering

To enhance predictive performance, new derived features were created: Shot Accuracy: Ratio of shots on target to total shots. Involvement Rate: Minutes played divided by total match duration (to assess influence). Pass Efficiency: Passes completed per minute. Disciplinary Weight: A weighted score combining yellow and red cards

For the xG regression model, we focused on attacking metrics such as shots, shots on target, assists, minutes played, and player position. For the MotM classification model, we included a broader range of features, including defensive metrics like tackles and cards, as well as passing statistics.

Modeling Approaches

Two types of models were developed using Python's Scikit-learn library:

1. Expected Goals (xG) – Regression Model

The treated the number of goals scored as a continuous target variable to model. Although traditional xG models use shot location and body part data, our approximation leverages available match-level statistics.

Models evaluated: a. Linear Regression, b. Random Forest Regression, c. Ridge Regression.

2. Man of the Match (MotM) – Classification Model

This is a binary classification problem where the target is 1 if the player was named MotM, 0 otherwise.

Models evaluated: a. Logistic Regression, b. Random Forest Classifier, c. Support Vector Machine (SVM).

Model Training and Evaluation

We split the data into training and testing sets using an 80:20 ratio. K-fold cross-validation (with $k = 5$) was used to ensure that the models generalize well to unseen data. For the regression task, evaluation metrics included Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 score. For classification, we used Accuracy, Precision, Recall, and F1-Score. Hyper parameters for each model were optimized using Grid Search Cross-Validation to improve performance. In both modeling tasks, feature importance was analyzed to interpret which metrics had the greatest influence on predictions.

Tools and Environment

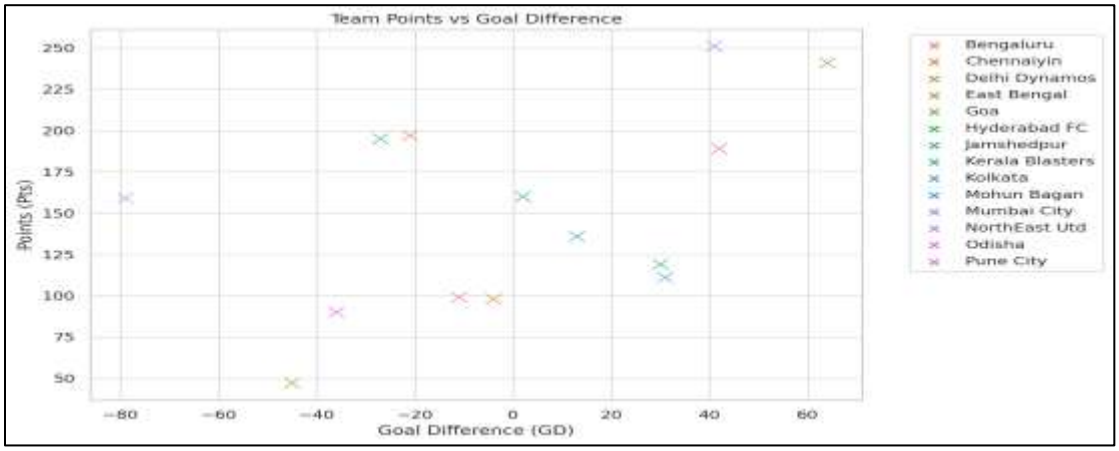
All processing and modeling were conducted in Python using libraries such as Pandas, NumPy, Scikit-learn, and Matplotlib for visualizations. The Jupyter Notebook environment was used for an interactive and iterative workflow.

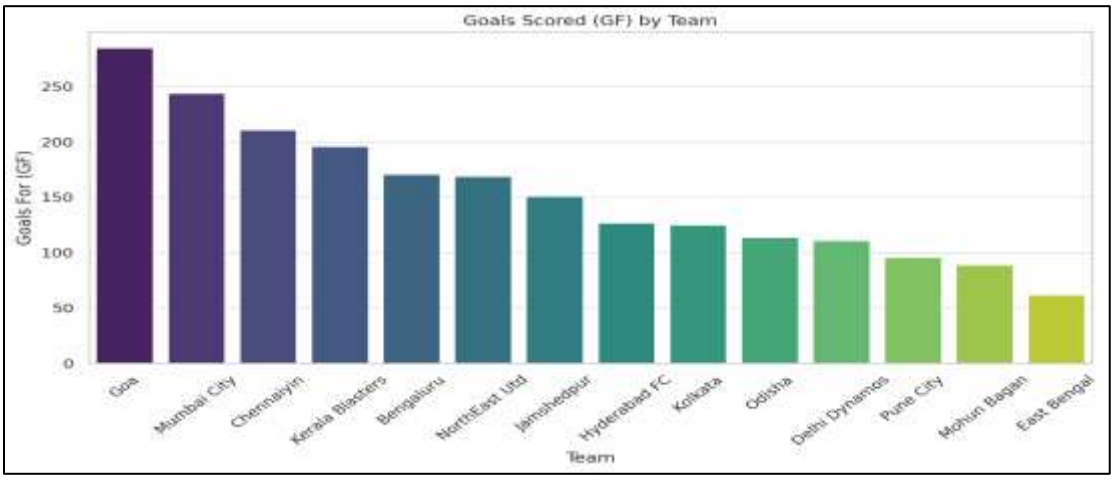
RESULT AND DISCUSSION

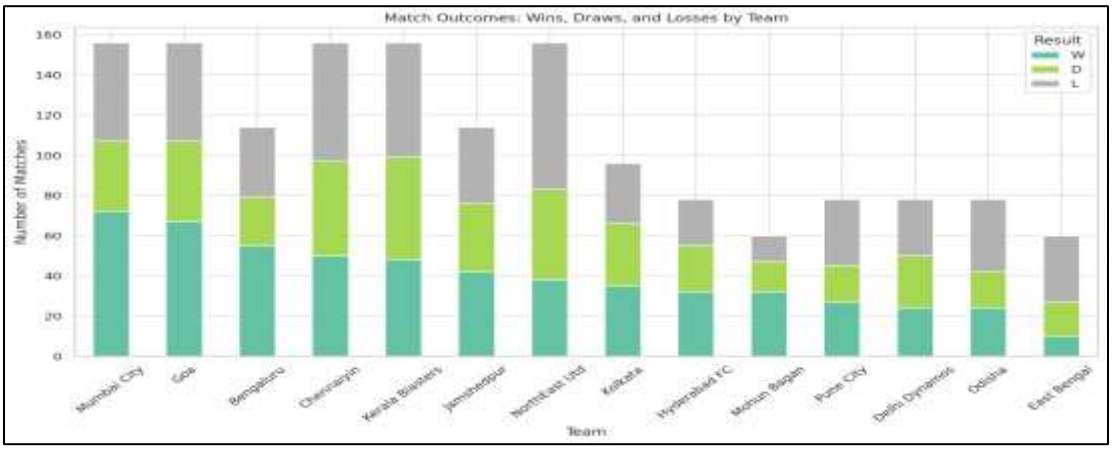
The Indian Super League (ISL) has evolved into a competitive and fast-paced football league since its inception. Though this study initially aimed at individual player performance modeling, the available data provided an opportunity to analyze team-level performance trends over the last decade. The results highlight key patterns in team dominance, scoring efficiency, and match outcomes, offering valuable insights into the league's competitive dynamics.

One of the most telling relationships visualized is between goal difference (GD) and total points (Pts). Teams that maintained a high positive goal difference tended to accumulate more points consistently. This is expected in football, where goal difference often mirrors both offensive strength and defensive solidity. Teams such as Mumbai City FC, Goa, and Bengaluru FC are seen leading this space, underlining their consistency in scoring more while conceding less. Interestingly, there are instances where teams with similar goal differences show noticeable variation in their total points, suggesting a possible difference in game management or match outcomes (e.g., winning vs. drawing games).

The distribution of Goals for (GF) further reveals which clubs have historically taken a more attacking approach. Goa stands out as the most aggressive team in terms of goal scoring, followed closely by Mumbai City FC. These clubs have been home to some of the league's most dynamic attacking players, which could explain their high scoring tallies. On the other hand, teams like North East United and Odisha FC have struggled to match this level of offensive output, correlating with their lower point totals.







The stacked bar chart showing wins, draws, and losses adds more contexts. While strong teams like Goa and Mumbai City boast a higher number of wins, other clubs have a more balanced

or loss-heavy distribution. This chart allows us to see which clubs play for wins versus those that frequently settle for draws or face losses. For instance, Hyderabad FC and Kerala Blasters show a moderate number of draws and losses, hinting at potential inconsistencies in converting matches into victories.

These team-level statistics point to a few key insights:
Offensive metrics (like goals scored and GD) strongly correlate with overall performance. Successful teams manage to win rather than draw close games, which is reflected in their higher point accumulation. There is room for performance optimization among lower-ranked teams through better game strategy, squad depth, or data-driven decision-making. Although this analysis is limited to aggregated team-level data, it still demonstrates the power of even basic statistics in drawing meaningful conclusions about performance. We can see the difference the playing abilities among teams, which we can find some more things apart from the competition result, it is very important of players physical capacities and strength to use their maximum potential (Panackal et al., 2012; Abraham, 2014); and the cardiovascular endurance is also an important factor to play 90 minutes (Kumar et al., 2025; Ashokan & Abraham., 2015). With more granular match-by-match or player-level data in the future, these insights can be made even richer and more actionable for coaches, analysts, and fans alike.

CONCLUSION

This study explored the application of data analytics and machine learning to evaluate team performance in the Indian Super League (ISL), using publicly available cumulative data. While our initial objective was to model individual player performance metrics such as expected goals (xG) and the likelihood of being named 'Man of the Match', the available dataset constrained the analysis to team-level metrics. Nevertheless, the insights gained from analyzing team performance were both meaningful and revealing.
Through visual exploration and statistical interpretation, we demonstrated that key indicators such as goal difference, goals scored, and match outcomes (wins, draws, losses) are closely tied to overall team success in the ISL. The consistent performance of teams like Goa, Bengaluru FC, and Mumbai City FC reflects a strong balance between attack and defense, while other teams reveal opportunities for improvement through strategic changes or data-informed decision-making.
This study findings underscore that even relatively high-level aggregate data, when analyzed systematically, can offer valuable insights into performance trends and competitiveness in football. This kind of analysis can be a foundation for more advanced work that benefits not only fans and analysts but also coaches and decision-makers within Indian football clubs.
In conclusion, this study serves as a foundational step toward more data-driven analysis in Indian football. With improved data availability and institutional support, Indian clubs can increasingly leverage analytics to inform tactics, training, and recruitment aligning with global best practices in modern football analytics.

Implication
There are several avenues for extending and refining this research:
1. Player-Level Data Integration: Accessing match-by-match player data, including granular actions like passes, tackles, shot locations, and positioning, would allow us to revisit the initial

goal of predictive modeling for individual players. This would unlock capabilities such as xG modeling; pass network analysis, and role-based player evaluation.
2. Temporal and Seasonal Analysis: Expanding the dataset to analyze season-over-season trends would help identify long-term strategic shifts, team evolution, and the impact of transfers or coaching changes.
3. Incorporation of Advanced ML Techniques: More sophisticated models like XGBoost, Neural Networks, or clustering algorithms could uncover hidden patterns in team strategies or player roles, especially when more detailed data is available.
4. Real-time Analytics and Visualization Dashboards: Building interactive tools or dashboards can make such analysis more accessible to ISL fans and professionals, enabling real-time tracking of team and player performance.

REFERENCES

1. Ashokan. K., & George Abraham. (2015). Influence of different intensity aerobic dance on BMI among young boys. International Journal of Physical and Social Sciences, 5 (10), pp. 591-598.
2. Bunker, R., & Thabtah, F. (2019). A machine learning framework for sport result prediction. Applied Computing and Informatics, 15(1), 27–33.
3. Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1851–1861.
4. George Abraham. (2010). Effect of progressive weight training on thigh girth, Journal of physical education and allied disciplines, 1(2), pp.97-99.
5. Abraham, G. (2014). Impact of moderate velocity resistance training on explosive strength among adolescents. International Journal of Fitness, Health, Physical Education & Iron Games, 1 (1), pp. 268-270.
6. Lucey, P., Bialkowski, A., Monfort, M., Carr, P., & Matthews, I. (2015). Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. MIT Sloan Sports Analytics Conference.
7. Martin Babu Panackal., Tony Daniel., & George Abraham. (2012). Effects of different training methods on power output among school team players, International journal of advanced scientific and technical research, 2 (5), pp.56-63.
8. Miller, A., Bornn, L., Adams, R. P., & Goldsberry, K. (2014). Factorized point process intensities: A spatial analysis of professional basketball. International Conference on Machine Learning, 235–243. (Referenced for cross-sport methodology comparison.)
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
10. Sæbo, H., & Hvattum, L. M. (2019). Using match event data and player ratings to predict football results. IMA Journal of Management Mathematics, 30(2), 159–176.
11. Sankaranarayanan, P. S., George Abraham., & Sajeev Jos. (2011). Eight weeks of different resistance training modes on big muscle hypertrophy of adolescents, Journal of experimental sciences, 2(2), pp.49-51.
12. Sarita Tyagi., Anil K. Vanaik., Monika Wasuja., Ranvir Singh., Surekha SudamDaptare., Sebastian K. M., Anu Joshi., & George Abraham. (2025). Exploring Circuit Training With

Different Techniques To Boost Anaerobic Capacity In Teens: An In- Depth Analysis And Collection Of Insights. Journal of Applied Bioanalysis, 11 (8), p. 25-30
13. Sathees Thomas., Sebastian K. M., Surekha Sudam Daptare., Toy C. T., Suman Pandey Mahadevan., Jijo Mathai., Chinsu Joy., & George Abraham. (2024). Diving Into How Different Competitive Sports Shape Motivational Levels For Athletes: A Comparative Analysis. Educational Administration: Theory and Practice, 30(4), 11612-11617.
14. Scikit learn Developers (2023). Scikit-learn: Machine Learning in Python. Retrieved from scikit-learn.org.
15. Suresh, S. (2023). Indian Super League Statistics 2014–2023 [Data set]. Kaggle. Retrieved from, www.kaggle.com/datasets/surajsuresh29/indian-super-league-statistics-2014-2023.
16. T. Frank Sunil Justus., & George Abraham. (2021). Stress among sports team captains at university level in Tamil Nadu. Adarsh-Journal of Management Studies, 13(1), pp. 36-46.
17. V. Sai Abhinav & George Abraham. (2022). Screening variant levels of mood disorders symptoms among interuniversity football players. Journal of Engineering, Computing & Architecture, 12(1), pp. 1-4.
18. Yogesh Kumar., Suma Joseph., Thushara Philip., Kishore Mukhopadhyay., K. Ketheeswaran., Dhirendra Kumar Singh., Hrishikesh Gopal., Jian Abdullah Noori., & George Abraham. (2025). Doing Aerobic Workouts And Elastic Strength Training Is A Great Way For College Novice Athletes To Enhance Their Cardiorespiratory Stamina. Afr. J. Biomed. Res. 27 (3), 3227-3232.