A Machine Learning Framework for Data-Scarce Regression using SMOGN with Joint Hyperparameter Optimization: A Case Study with Cricket Performance Prediction

Harthik Manichandra Vanumu*
Department of Computer Science and Engineering
Manipal Institute of Technology Bengaluru
Manipal Academy of Higher Education, Manipal, India
harthik.mitblr2023@learner.manipal.edu

Paranjay Lokesh Chaudhary*
Department of Computer Science and Engineering
Manipal Institute of Technology Bengaluru
Manipal Academy of Higher Education, Manipal, India
paranjay.mitblr2023@learner.manipal.edu

Usha Moorthy
Department of Information Technology
Manipal Institute of Technology Bengaluru
Manipal Academy of Higher Education, Manipal, India
m.usha@manipal.edu

Syed Anwar Ali
Department of Physical Education and Sports
Manipal Academy of Higher Education, Manipal, India
syed.s@manipal.edu

***Abstract*—This study presents a machine learning framework designed to enhance predictive accuracy in data-scarce regression tasks, a common challenge in predictive modeling. The core methodology integrates feature engineering, SMOGN data augmentation, and joint Optuna-based hyperparameter optimization of both SMOGN and multiple machine learning models. Evaluated using K-fold cross-validation across multiple random seeds, the framework demonstrated its effectiveness in predicting season-long batting performance in the Women's Premier League (WPL), an emerging cricket league with limited data serving as a demanding testbed. The results confirmed that SMOGN, when tuned jointly with model parameters, significantly improved predictive accuracy and stability, especially in ensemble models. An ablation study that excluded SMOGN quantified its impact, demonstrating that the impact of data augmentation varied across models, particularly for certain ensemble models that showed significant improvements. This research establishes a practical and adaptable machine learning framework with the potential to generate robust predictive models in various data-limited regression contexts. By integrating feature engineering, tailored data augmentation with efficient joint hyperparameter optimization, and comprehensive multi-seed evaluation, the framework demonstrated effectiveness in the WPL environment, thereby highlighting its potential value for various domains facing similar data constraints.**

***Index Terms*—Machine Learning, SMOGN, Joint Hyperparameter Optimization, Data-Scarce Regression, Sports Analytics**

I. Introduction

The development of predictive models in data-scarce regression contexts is a common challenge encountered wherever historical data is limited. This issue is prevalent in various domains, such as finance, healthcare, and sports analytics, with possible data-scarce scenarios, including early-stage financial forecasting, the prediction of rare disease outcomes, and specialized engineering problems. Such constraints can undermine the reliability of conventional machine learning approaches. Sports analytics is also a field that often faces similar data limitations. The recent launch of the Women's Premier League (WPL), by the Board of Control for Cricket in India (BCCI) in 2023 and modeled after the successful men's Indian Premier League (IPL), serves as an illustrative example. Due to its recent establishment, the WPL inherently has limited historical data. This scarcity complicates the prediction of key performance indicators such as season-long player batting performance, thus creating a valuable and demanding testbed for novel methodologies.

The application of machine learning in cricket, particularly in established leagues, has been extensive, with studies focusing on predicting match outcomes, individual player scores, and various performance metrics [1]–[6]. A key challenge in new leagues like the WPL is the lack of extensive historical data, which hampers the development of reliable predictive models. Although some research has explored women's cricket performance [7], [8] or specific aspects of T20 analytics [9], a gap remains in methodologies specifically designed to predict comprehensive season-long batting performance (e.g. total runs scored) within the severe data constraints characteristic of a new league like the WPL [10]. Moreover, the application of advanced data augmentation techniques tailored for regression tasks and their efficient hyperparameter optimization in conjunction with model training remains relatively underexplored in this specialized area of sports analytics.

*These authors contributed equally to this work.

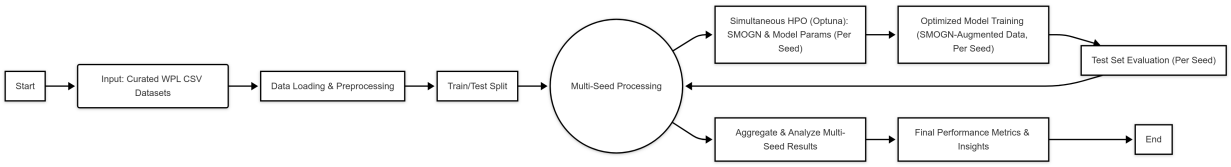


Fig. 1: Data Processing and Modeling Workflow.

To address this gap, we introduce a machine learning framework, designed for broader applicability in data-scarce regression, which we demonstrate by predicting season-long batting performance in data-scarce environments, using the WPL as a case study. The proposed framework integrates feature engineering (including player valuation, opportunity, and form proxies) with SMOGN (Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise) [11] data augmentation. A key contribution of this work is the joint hyperparameter optimization of the SMOGN algorithm and nine diverse regression models using Optuna [12]. This approach, tailored for data-scarce regression tasks, represents a novel application for enhancing predictive accuracy under such constraints.

To demonstrate its practical value, the framework is comprehensively evaluated on the WPL dataset. This evaluation involved testing its predictive performance through cross-validation across multiple seeds and included an ablation study to understand the contribution of SMOGN data augmentation, which showed it worked well for certain models.

The remainder of this paper is structured as follows. Section II details the methodology, including data acquisition, preprocessing, feature engineering, the SMOGN augmentation technique with joint optimization, and the experimental setup. Section III presents and analyzes the empirical results, including the performance of various models and the findings of the ablation study. Section IV discusses the limitations of the study, and Section V concludes the paper, summarizing the key findings and contributions. Section VI suggests avenues for future research.

II. Methodology

This section describes the comprehensive machine learning pipeline developed to predict season-long batting performance ("runs" scored) in the data-scarce context of the Women's Premier League (WPL). The workflow encompasses data acquisition and preparation, feature engineering, the core SMOGN data augmentation strategy with joint hyperparameter optimization, model training, and the multi-seed evaluation.

A. Overview

Building on prior cricket analytics research that targets "runs" as a key performance indicator, this study predicts the total runs scored by a player in a WPL season. This season-level prediction target was selected due to the limited dataset size available from the initial WPL seasons (approximately 109 unique player-season instances after cleaning and before augmentation, compiled from three CSV files). This target provides a holistic measure of a player's batting contribution, which is relevant for seasonal evaluations, scouting, and strategic planning. However, it presents inherent challenges due to data sparsity compared to match-level predictions. A batting-centric approach was adopted, utilising batting-related metrics to predict the "runs" variable. The overall data processing and modeling pipeline is illustrated in Fig. 1.

B. Data Acquisition and Preparation

The dataset for this study was derived from ball-by-ball records for three Women's Premier League (WPL) seasons (2023-2025), sourced primarily from Cricsheet [13] and supplemented with scraped data from ESPNcricinfo [14] and the official WPL website[1]. This raw data underwent extensive preprocessing, which included YAML parsing, data aggregation, validation, integration of supplementary information like player auction prices ('price'), and manual checks to ensure data consistency, ultimately yielding three curated CSV files, each containing data for one WPL season. These CSVs served as the direct "Input: Curated WPL CSV Datasets" for the machine learning pipeline (Fig. 1). Within this pipeline, the CSVs were loaded and merged, followed by essential preparation steps such as feature and target column validation, numeric type conversion, and a run-specific seeded shuffle to ensure unbiased data partitioning for modeling.

C. Data Cleaning And Preprocessing

Before training the model, the prepared dataset was subjected to a standard preprocessing sequence using tools from the scikit-learn library [15]. Applied exclusively to the training data to prevent data leakage, this preprocessing pipeline involved the following steps:

1) Imputation: Missing values, which can arise from data collection or the numeric conversion process, were addressed through a median imputation strategy. This approach was selected for its robustness to outliers.
2) Scaling: All features were subsequently standardized. This process transforms features to have zero mean and unit variance, a common requirement for models sensitive to the scale of input features, such as linear models.
3) Transformation: To mitigate skewness in the distributions of the features and to approximate a more normal distribution, a Yeo-Johnson power transformation was applied.

[1]Data validated using the official Women's Premier League website operated by the BCCI: https://www.wplt20.com (Accessed: 2025-05-27).

D. Feature Selection and Engineering

The feature set was designed to capture various factors that influence a player's seasonal run output, including their valuation, opportunities, and performance characteristics. The raw data sources for the input CSVs were assumed to be processed from match-level Key Performance Indicators (KPIs) commonly used in T20 cricket analytics. The final feature set used for prediction in this study comprised the following eight variables: price (player auction price), role (player category), balls_faced (opportunity proxy), and precalculated weighted average metrics, including weighted_avg_sr, weighted_avg_rbw, weighted_avg_hard_hitter, weighted_avg_dot_percent, and weighted_avg_bound_percent. These features were designed to capture player valuation, opportunity, scoring efficiency, power hitting, and scoring patterns. The specific calculation of the weighted average metrics was external to the described pipeline. These external calculations were specifically designed to prevent data leakage from test or future data in the features used for training.

E. Data Augmentation using SMOGN with Joint Hyperparameter Optimization

To address the challenge of limited training data, SMOGN (Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise) was applied as the primary data augmentation technique. SMOGN enhances model generalization by generating synthetic data points, particularly for underrepresented values in the target variable's ('runs') distribution. This augmentation was applied exclusively to the preprocessed training data. A key element of this framework is the joint hyperparameter optimization of both the SMOGN algorithm and the selected machine learning models. This joint approach is crucial because the optimal augmentation strategy heavily depends on the chosen model and its parameters, and vice versa, an interaction that sequential optimization might miss. This was achieved using the Optuna optimization framework, a tool to automate hyperparameter search processes. 50 Optuna trials were conducted for each model and seed. The number of trials was determined empirically, aiming to balance a robust exploration of the hyperparameter space with computational feasibility. In every trial, Optuna concurrently proposed a complete set of hyperparameters for both the SMOGN algorithm and the regression model itself. The SMOGN hyperparameters included k neighbours, relevance threshold, relevance coefficient, and sampling method. The configured model was then trained and evaluated on this augmented data using 3-fold cross-validation (KFold with shuffling and a fixed random seed to ensure consistency in the data splits), with the mean R^2 score serving as the optimization objective. The Mean Squared Error (MSE) was also tracked during this process.

F. Data Partitioning, Model Training, and Stability Analysis

The preprocessed dataset was split into a 80% training portion and a 20% test portion, using a run-specific random seed to ensure consistent partitioning across experiments.

TABLE I: Machine Learning Regression Models Evaluated

S. No	Name of the Model used
1	AdaBoost
2	CatBoost
3	Extra Trees
4	Gradient Boosting
5	Lasso Regression
6	Linear Regression
7	Random Forest
8	Ridge Regression
9	XGBoost

Nine different regression models (detailed in Table I) were selected for evaluation, chosen to represent a spectrum of algorithmic families. These include linear approaches (Lasso Regression, Linear Regression, Ridge Regression) and non-linear approaches comprising tree-based ensembles that encompass bagging techniques (Random Forest, Extra Trees) and boosting techniques (AdaBoost, CatBoost, Gradient Boosting, XGBoost), thereby facilitating a comprehensive assessment of the framework's impact across different learning strategies.

Each model was subjected to the joint SMOGN and hyperparameter optimization protocol described in Section II.E. Upon completion of the Optuna optimization phase, the best-identified combination of SMOGN and model hyperparameters was used. SMOGN augmentation, using these optimal parameters, was applied once more to the entire preprocessed training set. The model, configured with its optimal hyperparameters, was then retrained on this fully augmented training dataset and subsequently evaluated on the unseen test set to assess its final performance. To ensure the reliability and stability of the reported results, the entire experimental procedure, from data partitioning to final model evaluation, was replicated five times. Each replication used a different global random seed ([42, 123, 456, 789, 1011]). The performance metrics presented in Section III are the mean and standard deviation derived from these five independent runs.

G. Ablation Study: Evaluation without SMOGN Data Augmentation

To rigorously evaluate the contribution of SMOGN data augmentation to model performance, an ablation study was conducted. This study mirrored the primary experimental setup in terms of data sources, preprocessing, the set of nine models, the 5-seed multi-run evaluation, and the use of Optuna for model hyperparameter optimization (50 trials per model). The key distinction was the exclusion of the SMOGN augmentation step. In this ablation study, the models were trained and tuned directly on the original preprocessed training data without any synthetic sample generation. The results of this 'No SMOGN' pipeline provide a direct baseline for evaluating the efficacy of the SMOGN augmentation detailed in this framework.

III. Results and Discussion

This section presents an empirical evaluation of the proposed machine learning framework. First, it analyzes the performance of the nine regression models trained with the joint

TABLE II: Aggregated Model Performance Metrics (Mean ± Std Dev across 5 Seeds With SMOGN Augmentation and joint hyperparameter optimization)

Model	Train R^2 (Mean ± Std)	Test R^2 (Mean ± Std)	Train MSE (Mean ± Std)	Test MSE (Mean ± Std)	Mean CV R^2 (Mean ± Std)	Mean CV MSE (Mean ± Std)
Random Forest	0.9887 ± 0.0011	0.9355 ± 0.0338	60.0416 ± 8.0456	408.2295 ± 245.4474	0.8826 ± 0.0175	590.0696 ± 117.7476
AdaBoost	0.9719 ± 0.0030	0.9341 ± 0.0235	149.3230 ± 19.5712	403.8994 ± 181.9230	0.8701 ± 0.0247	648.7096 ± 147.8890
ExtraTrees	1.0000 ± 0.0000	0.9314 ± 0.0403	0.0000 ± 0.0000	437.6102 ± 284.0216	0.9077 ± 0.0097	465.8166 ± 88.1509
Gradient Boosting	0.9987 ± 0.0003	0.9180 ± 0.0431	6.9609 ± 1.9181	515.8611 ± 313.2442	0.8831 ± 0.0126	593.1930 ± 118.2421
XGBoost	1.0000 ± 0.0000	0.9170 ± 0.0358	0.0000 ± 0.0000	528.3567 ± 286.4671	0.8763 ± 0.0157	625.6408 ± 101.1535
Lasso	0.8969 ± 0.0189	0.8778 ± 0.0376	546.2016 ± 96.1211	754.7396 ± 323.5959	0.8668 ± 0.0079	656.1709 ± 59.7163
Ridge	0.8983 ± 0.0168	0.8699 ± 0.0357	536.5659 ± 94.9482	800.4116 ± 319.8485	0.8603 ± 0.0101	688.9396 ± 53.4372
LinearRegression	0.8987 ± 0.0189	0.8644 ± 0.0350	536.3828 ± 95.0274	825.5125 ± 302.9841	0.8554 ± 0.0118	767.6212 ± 62.4612
CatBoost	0.9999 ± 0.0000	0.8501 ± 0.0922	0.4436 ± 0.0703	951.3979 ± 667.3517	0.8310 ± 0.0302	890.5973 ± 238.1121

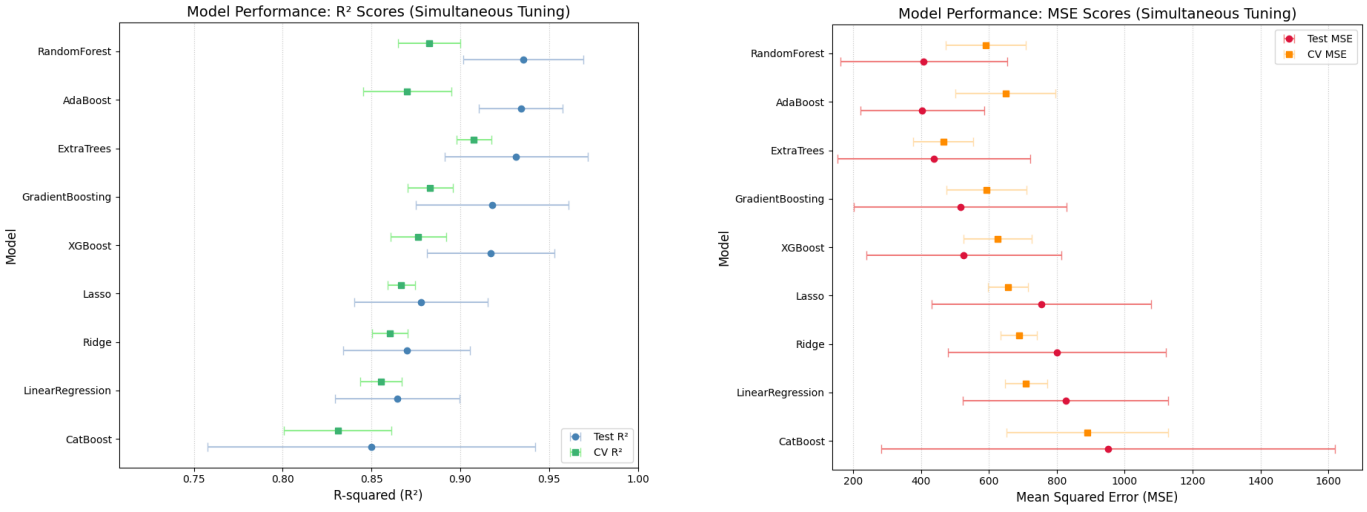
SMOGN and hyperparameter optimization strategy. Following this, the section presents the findings of the ablation study, which quantified the specific impact of SMOGN augmentation. All results are reported as the mean ± standard deviation across the five independent runs using different random seeds.

A. *Performance of Models with SMOGN Augmentation*

The aggregated performance metrics for the nine machine learning models, following joint SMOGN and optimization of the model hyperparameters using Optuna, are summarized in Table II. These results illustrate the models' predictive capabilities on the held-out test set, as well as their consistency during cross-validation within the training phase. Fig. 2(a) provides a visual comparison of the mean scores of the Test R^2 and the mean cross-validation (CV) R^2 across the models, while Fig. 2(b) provides a similar comparison for the Test MSE and CV MSE, offering a graphical representation of performance and stability.

As shown in Table II and in Fig. 2(a) and Fig. 2(b), the ensemble methods demonstrated strong predictive capabilities when SMOGN augmentation was jointly optimized with the model. Random Forest achieved the highest mean Test R^2 of approximately 0.9355 ± 0.0338, followed closely by AdaBoost with a mean Test R^2 of 0.9341 ± 0.0235. AdaBoost also produced the lowest mean Test MSE among these top performers, approximately 403.90 ± 181.92. This pattern is also evident when comparing the error magnitudes in Fig. 2(b). Extra Trees also performed robustly, with a mean Test R^2 of 0.9314 ± 0.0403. The standard deviations for Test R^2 in these leading models, as illustrated by the error bars in Fig. 2(a), were relatively low (ranging from 0.0235 to 0.0403), suggesting good stability between the different experimental runs. The average Test MSE for AdaBoost translates to a Root Mean Squared Error (RMSE) of approximately 20.1 runs, indicating that the model's predictions are typically within this range of the actual season total, offering a valuable baseline for player evaluation.

The training performance metrics, particularly the very high Train R^2 scores (often approaching 1.0) and low Train MSE values for models such as ExtraTrees, XGBoost, and CatBoost, suggest potential overfitting to the augmented training data. This is not uncommon when using complex models with aug-



(a) R^2 Scores Comparison.

(b) MSE Scores Comparison.

Fig. 2: Comparison of Mean Test vs. Cross-Validation R^2 Scores (a) and MSE Scores (b) (± Std Dev across 5 Seeds) Across Models (With SMOGN Augmentation).

mented datasets. However, the strong Test R^2 scores (generally above 0.91 for several ensemble methods, as seen in Fig. 2(a)) confirm the effective generalizability to unseen data within this framework. Simpler linear models (Lasso, Ridge, Linear Regression) exhibited more modest test performance compared to the top ensemble techniques when SMOGN was applied, a distinction clear in both Fig. 2(a) and Fig. 2(b). CatBoost, despite high training scores, showed lower test efficacy and higher variability in this augmented setting. These results highlight the effectiveness of boosting and bagging ensembles for effectively learning from synthetically augmented data in data-scarce regression tasks, showcasing their ability to model complex relationships within such enhanced feature spaces.

B. Ablation Study: Impact of SMOGN Augmentation

An ablation study was conducted to evaluate the contribution of SMOGN data augmentation and its joint optimization. In this study, the models were trained and tuned according to the procedures outlined in Section II.G, excluding any data augmentation. The performance of key models from this 'No SMOGN' pipeline was compared against the results of the primary 'with SMOGN' pipeline (Table II, Fig. 2(a), Fig. 2(b). Without SMOGN, Gradient Boosting was the model that performed the best, achieving a mean Test R^2 of 0.9009 ± 0.0232 and a Test MSE of 393.65 ± 57.74. CatBoost followed with a Test R^2 of 0.8976 ± 0.0195 and Test MSE of 413.55 ± 95.46. For Random Forest, the 'No SMOGN' pipeline yielded a Test R^2 of 0.8842 ± 0.0423 and Test MSE of 443.26 ± 69.69 while AdaBoost achieved a Test R^2 of 0.8814 ± 0.0320 and Test MSE of 470.48 ± 86.73.

A comparison of these outcomes indicates that the application of SMOGN with joint optimization significantly improved performance across several models. Random Forest's Test R^2 improved from ≈ 0.8842 (No SMOGN) to ≈ 0.9355 (With SMOGN, Table II, as illustrated in Fig. 2(a), and its Test MSE decreased from ≈ 443.26 to ≈ 408.23 (as seen by comparing with Fig. 2(b). Similarly, AdaBoost's Test R^2 increased from ≈ 0.8814 to ≈ 0.9341 (Fig. 2(a), with its Test MSE reducing from $\approx$ 470.48 to $\approx$ 403.90 (Fig. 2(b). Extra Trees also showed a marked improvement in Test R^2 with augmentation (from ≈ 0.8963 without SMOGN to ≈ 0.9314 with SMOGN, Table II, Fig. 2(a).

However, the impact of SMOGN was not universally beneficial across all algorithms under this optimization strategy. For Gradient Boosting, while the Test R^2 slightly increased with SMOGN (from $\approx$ 0.9009 to $\approx$ 0.9180, Table II, Fig. 2(a), this came with a considerable increase in Test MSE (from $\approx$ 393.65 to $\approx$ 515.86, Fig. 2(b) and higher variability. CatBoost's performance, in terms of Test R^2, declined with SMOGN compared to its unaugmented counterpart (from $\approx$ 0.8976 without SMOGN to $\approx$ 0.8501 with SMOGN, Table II, Fig. 2(a).

This ablation study highlights that SMOGN, when jointly optimized with model hyperparameters, significantly enhances the predictive accuracy of ensemble methods such as Random Forest and AdaBoost in data-scarce scenarios. However, SMOGN's effectiveness remains dependent on the model. The greater benefit of SMOGN for these tree-based ensembles might stem from their ability to capture complex nonlinear interactions within the augmented feature space, effectively utilizing the synthetic samples. In contrast, Gradient Boosting showed an increase in R^2 but with a higher MSE and variability with SMOGN. It may have been more sensitive to specific characteristics or potential noise introduced by the synthetic samples when it was set up in the way previously found to be best for operating without data augmentation, thereby highlighting a complex interaction. The findings validate SMOGN as a beneficial component of the proposed framework for specific algorithms, while also highlighting the necessity of comparative analysis to determine the optimal approach for each type of model.

IV. Limitations

The findings of this study are subject to certain limitations. The primary constraint is the reliance of the dataset on only three WPL seasons, which, despite SMOGN augmentation, can affect the generalizability of the model and contribute to observed run-to-run variability. Furthermore, the prediction scope was restricted to batting-specific features and pre-calculated metrics, excluding key game factors such as opposition details or venue data, thus limiting the model's breadth. Finally, although the 50-trial Optuna search for joint hyperparameter optimization was systematic, it may not have fully explored the extensive combined parameter space required to achieve the absolute global optimum.

V. Conclusion

This study introduces a machine learning framework for predicting season-long batting performance ('runs' scored) in the data-scarce Women's Premier League. A key contribution of this work is the integration of feature engineering with SMOGN data augmentation, where the parameters for both SMOGN and nine diverse regression models were jointly optimized using the Optuna framework. Rigorous multi-seed evaluation demonstrated that this integrated approach enabled ensemble models, notably Random Forest and AdaBoost, to achieve high predictive accuracy and stability (e.g., mean Test $R^2 \approx 0.9355$ for Random Forest). The findings of the ablation study revealed that the jointly tuned SMOGN augmentation provides significant benefits, although these depended on the model. This research provides two main contributions. First, it presents a practical and adaptable machine learning framework that, through its use of advanced data augmentation and efficient optimization strategies validated in the WPL, demonstrates strong potential for application in various data-limited regression domains such as financial forecasting or healthcare outcome prediction and any domain facing similar constraints for regression tasks. Second, it establishes a methodology that provides specific and actionable predictive insights to support decisions in diverse sporting contexts, including leagues, national teams, and tournaments. This is particularly relevant for applications such as player scouting

and performance management demonstrating efficacy in data-scarce environments such as the WPL.

VI. Future Work

Future research can extend this work across multiple avenues. A key consideration is improving generalizability by validating the machine learning framework across multiple settings where regression is applied, particularly when data is limited. This could involve exploring its use in areas such as financial forecasting or healthcare outcomes prediction without being restricted to these specific examples. Furthermore, testing its application on larger datasets from diverse sporting contexts, including future WPL seasons, will further assess its robustness as domain-specific dynamics evolve. Methodological advancements will include exploring improved techniques to measure a player's current performance trend and incorporating more specific game aspects such as venue characteristics or opposition bowling metrics [16]. Integrating Automated Machine Learning (AutoML) techniques [17] presents a promising direction for optimizing the entire pipeline, including feature selection, model selection, and joint optimization of model and data augmentation parameters. Further exploration of alternative data augmentation techniques specifically suited for data-scarce regression scenarios along with investigations into deep learning architectures, particularly in cases where more complex or sequential data becomes available could enhance predictive capabilities. Furthermore, enhancing model interpretability through Explainable AI (XAI) methods [18] may offer deeper insights into the key factors that influence predictions, offering practical applications that extend beyond predictive accuracy.

References

[1] M. Sumathi, S. Prabu, and M. Rajkamal, "Cricket players performance prediction and evaluation using machine learning algorithms," in *Proceedings of the 2023 International Conference on Networking and Communications (ICNWC)*, Apr. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10127503

[2] Y. Agrawal and K. Kandhway, "Winner prediction in an ongoing one day international cricket match," *Journal of Sports Analytics*, vol. 9, pp. 305–318, Feb. 2024. [Online]. Available: https://journals.sagepub.com/doi/epub/10.3233/JSA-220735

[3] P. Singh, J. Kaur, and L. Singh, "Predicting ipl victories: An ensemble modeling approach using comprehensive dataset analysis," in *Proceedings of the 2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications (AIMLA)*. IEEE, Mar. 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10531489

[4] Y. Jadwani, J. Denholm-Price, and G. Hunter, "A machine learning-based approach to analyse player performance in t20 cricket internationals," in *10th MathSport International Conference Proceedings 2023*, Budapest, Hungary, 2023. [Online]. Available: https://www.mathsportinternational.com/MathSport2023Proceedings.pdf

[5] D. J. Vestly, S. Hariharan, V. Kukreja, A. B. Prasad, K. Swaraj, and D. Gopichand, "Parametric analysis of a cricketer's performance using machine learning approach," in *Proceedings of the 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*. Madurai, India: IEEE, May 2023, pp. 344–348. [Online]. Available: https://doi.org/10.1109/ICICCS56967.2023.10142664

[6] S. Sanjaykumar, S. Natarajan, P. Y. Lakshmi, and F. A. Boby, "Predicting team success in the indian premier league cricket 2024 season using random forest analysis," *Physical Education Theory and Methodology*, vol. 24, no. 2, pp. 304–309, Apr. 2024. [Online]. Available: https://tmfv.com.ua/journal/article/view/2619/1706

[7] S. Mandoli, D. Sharma, and H. C. Joshi, "A discriminant model for skill oriented prediction of female cricketers depending upon selected performance parameters," *Physical Education Theory and Methodology*, vol. 21, no. 4, pp. 293–298, Dec. 2021. [Online]. Available: https://tmfv.com.ua/journal/article/view/1531/1435

[8] K. Gupta, "An integrated batting performance analytics model for women's cricket using principal component analysis and gini scores," *Decision Analytics Journal*, vol. 4, p. 100109, Sep. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772662222000467

[9] P. Puram, S. Roy, D. Srivastav, and A. Gurumurthy, "Understanding the effect of contextual factors and decision making on team performance in twenty20 cricket: an interpretable machine learning approach," *Annals of Operations Research*, vol. 325, no. 1, pp. 261–288, Jun. 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10479-022-05027-1

[10] P. Y. Lakshmi, S. Sanjaykumar, M. Dharuman, and A. Elangovan, "Using support vector regression kernel models for cricket performance prediction in the women's premier league 2024," *Physical Education Theory and Methodology*, vol. 24, no. 1, pp. 72–78, Feb. 2024. [Online]. Available: https://www.tmfv.com.ua/journal/article/view/2535/1678

[11] P. Branco, L. Torgo, and R. P. Ribeiro, "Smogn: a pre-processing approach for imbalanced regression," in *Proceedings of Machine Learning Research 74:36–50, 2017*, 2017. [Online]. Available: https://proceedings.mlr.press/v74/branco17a/branco17a.pdf

[12] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, Jul. 2019, pp. 2623–2631. [Online]. Available: https://doi.org/10.1145/3292500.3330701

[13] S. Rushe, "Cricsheet: Freely-available structured data for cricket," 2025, accessed: 2025-05-27. [Online]. Available: https://cricsheet.org/

[14] ESPNcricinfo, "Espncricinfo: Match schedules, results, and player data," https://www.espncricinfo.com, 2025, accessed: 2025-05-27.

[15] F. P. et al., "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Nov. 2011. [Online]. Available: https://dl.acm.org/doi/10.5555/1953048.2078195

[16] R. Lokhande, R. Awale, and R. Ingle, "Analysing the impact of field conditions, pitch features, and opponent strength on cricket performance: A machine learning approach," *VJSTE*, vol. 66, no. 3, pp. 3–14, Sep. 2024. [Online]. Available: https://vietnamscience.vjst.vn/index.php/vjste/article/view/1215/468

[17] M. Baratchi, C. Wang, S. Limmer, and et al., "Automated machine learning: past, present and future," *Artificial Intelligence Review*, vol. 57, no. 5, p. 122, Apr. 2024, accepted 10 February 2024; Published 18 April 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10462-024-10726-1

[18] P. R. Gogulamudi and Y. V. Pavan Kumar, "Explainable ai (xai): Explained," in *2023 IEEE International Conference on e-Stream and Emerging Technologies (e-STREAM)*. Hyderabad, India: IEEE, Apr. 2023, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10134984