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

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

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

Usha Moorthy
School of Computer Engineering
Manipal Institute of Technology Bengaluru
Manipal Academy of Higher Education, Manipal, India
m.usha@manipal.edu

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

***Abstract*—This study presents a machine learning framework to improve predictive accuracy in regression under data scarcity, a prevalent challenge in predictive modeling. A key contribution is a joint hyperparameter optimization strategy that integrates data augmentation with model training, outperforming traditional sequential approaches. Our approach simultaneously tunes SMOGN (Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise) and multiple regression models using Optuna, identifying optimal parameter combinations that sequential methods may miss. The framework was evaluated using k-fold cross-validation and multi-seed experiments on season-long batting performance prediction in the Women's Premier League (WPL), an emerging cricket league with limited historical data. The results show that tree-based ensembles consistently outperform linear models, with top performers achieving mean test R^2 values above 0.89. CatBoost achieved the highest mean test R^2 of 0.9075, with a standard deviation of 0.0162. An ablation study without SMOGN confirms the importance of this integrated augmentation strategy. By treating the pipeline as a fully integrated system, the framework provides a practical approach to predictive modeling under severe data constraints, with applicability across domains such as sports analytics, finance, and healthcare, and can serve as a blueprint for enhancing data synthesis within AutoML pipelines.**

***Index Terms*—Automated Machine Learning (AutoML), Hyperparameter Optimization, Data Augmentation, SMOGN, Sports Analytics**

I. Introduction

Predictive modeling in emerging domains, from early-stage financial forecasting to newly launched sports leagues, is often hindered by limited data, posing a significant challenge for AI and machine learning frameworks. Data scarcity amplifies imbalances, where rare but crucial outcomes are underrepresented, reducing predictive reliability and model robustness. These challenges require methodologies capable of delivering accurate predictions in low-data settings while ensuring generalization.

Conventional approaches to data scarcity, such as resampling, synthetic data generation, and transfer learning, provide partial solutions. However, most workflows treat data augmentation and model optimization independently, overlooking their potential synergy. Sequential pipelines often fail to leverage the interaction between augmentation and hyperparameter tuning, limiting predictive performance on tabular datasets that are small, diverse in structure, and skewed in distribution.

Sports analytics offers a practical and high-impact domain to test scarcity-aware regression methods. While established men's cricket leagues, like the Indian Premier League, benefit from extensive historical data, newly formed leagues such as the Women's Premier League (WPL) face low-data scenarios with limited historical records. This gap highlights the need for a joint optimization framework tailored for structured, data-scarce environments.

To address these challenges, we propose a general-purpose joint optimization framework that integrates SMOGN-based data augmentation with hyperparameter tuning across multiple regression models. By optimizing augmentation and model parameters simultaneously, the framework improves predictive accuracy and generalization in data-scarce environments. This integrated approach overcomes the limitations of sequential pipelines and provides a systematic methodology for scarcity-aware regression on structured tabular datasets.

We demonstrate the framework through the prediction of

*Both authors contributed equally to this work.

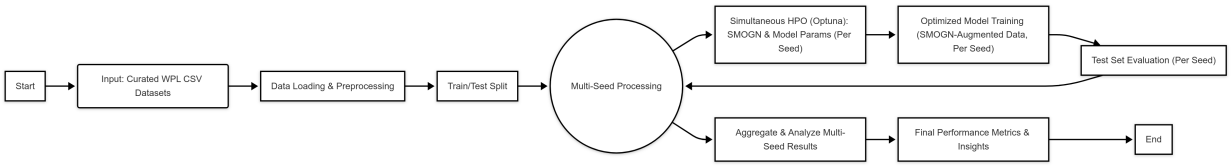


Fig. 1: The joint optimization framework for WPL batting performance prediction

season-long batting performance in the WPL, showing that it produces reliable and stable predictions despite limited data. Beyond sports, the methodology generalizes to other structured tabular domains, including finance, healthcare, and engineering, where historical data is scarce. The remainder of the paper is organized as follows. Section II reviews related work, Section III details the methodology, Section IV presents results, Section V discusses limitations, Section VI concludes, and Section VII outlines directions for future research.

II. Related Work

Data scarcity in regression has motivated numerous AI and ML approaches aimed at mitigating the underrepresentation of rare outcomes. Common strategies include synthetic oversampling, resampling, and transfer learning. In particular, SMOGN, a regression-specific oversampling technique, generates synthetic samples with Gaussian noise in underrepresented regions of the target variable. [1]. Hyperparameter optimization frameworks such as Optuna facilitate the efficient exploration of large model configurations [2]. However, most existing workflows apply data augmentation and model tuning sequentially, potentially limiting predictive performance in data-scarce conditions.

The effectiveness of data augmentation is highly domain-specific. Although augmentation techniques for image and text are well established, structured tabular datasets present unique challenges. They are often small, heterogeneous, and imbalanced, which can limit the efficacy of conventional augmentation strategies. Although explainability and interpretability are increasingly emphasized in decision-critical domains [3], they remain secondary considerations in scarcity-aware regression pipelines.

Sports analytics offers a practical, real-world testbed for structured tabular regression. In men's cricket leagues, ML has been widely applied to predict player performance [4], [5], team outcomes [6], and match winners [7], [8], conduct parametric analyses [9], and develop performance indices [10]. Other research has explored the effect of contextual factors on team performance [11]. However, these studies rely on abundant historical data. In contrast, women's cricket, particularly the recently launched WPL, presents a uniquely data-scarce environment. Prior studies in this area include discriminant models for role classification [12], PCA-based batter rankings [13], and kernel comparisons for regression [14], but have not focused on joint optimization of augmentation to overcome scarcity.

Collectively, these observations highlight two gaps: (i) most regression pipelines treat data augmentation and model tuning as independent steps, and (ii) systematic frameworks for scarcity-aware regression on small, structured tabular datasets remain rare. Our work addresses both gaps by integrating SMOGN-based augmentation with joint hyperparameter optimization.

III. 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 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. We adopted a batting-centric approach, using 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 [15] and supplemented with scraped data from ESPNcricinfo [16] 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

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

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 model training, we applied a standard preprocessing sequence to the dataset using the scikit-learn library [17]. To prevent data leakage, we fitted this pipeline exclusively on the training data. The process consisted of the following sequential steps:

1) Imputation: We used a median imputation strategy to fill in any missing values. We chose this method because of its robustness to outliers.
2) Scaling: We then standardized all features to have a zero mean and unit variance. This step ensures that models sensitive to the scale of input data, such as linear models, can perform effectively.
3) Transformation: Finally, we applied a Yeo-Johnson power transformation to mitigate skewness and normalize the feature distributions.

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 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. 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. The Joint Hyperparameter Optimization Framework

A primary contribution of this work is the proposed joint hyperparameter optimization framework, which simultaneously tunes data augmentation and model parameters as a unified system. This approach is more effective than traditional sequential methods. In a sequential approach, one first identifies the optimal SMOGN hyperparameters and then tunes the model on the resulting static augmented dataset. Such a strategy is suboptimal because it ignores the significant interaction between data augmentation and model configuration. Although sequential methods are commonly used in AutoML pipelines to reduce the computational burden of hyperparameter exploration, they may fail to identify globally optimal combinations. The optimal data augmentation strategy is dependent on the model, and conversely, the optimal model configuration is dependent on the augmented data.

Our framework addresses this challenge by treating data augmentation and model training as a single, integrated pipeline. To mitigate the limitations of limited training data, we use SMOGN (Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise) [1], which generates synthetic data points to enrich underrepresented regions of the target variable's distribution, thereby enhancing model generalization.

Hyperparameter optimization is performed using the Optuna framework [2], which simultaneously searches for the best combination of parameters for both the regression model and SMOGN. In each of 50 optimization trials, a complete set of hyperparameters, including SMOGN's number of neighbours, relevance threshold, relevance coefficient, and sampling method, is proposed and evaluated together with the model configuration.

Each trial is assessed using 5-fold cross-validation, with the mean R^2 score serving as the optimization objective. During cross-validation, SMOGN augmentation is applied to the training folds but not to the validation fold to prevent data leakage. The relevance function for SMOGN is computed based on the quantiles of the training data in each fold. This joint approach enables the discovery of globally superior configurations, where the augmentation strategy is tailored to the model, resulting in higher performance in data-limited environments.

To improve efficiency, two pruning strategies are employed: trials that generate invalid synthetic data are immediately pruned, and a median pruner terminates trials based on the mean cross-validated R^2 from previous folds.

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. Nine different regression models were selected for evaluation, chosen to represent a spectrum of algorithmic families. These include linear approaches (Lasso Regression, Linear Regression, Ridge Regression) and other approaches comprising tree-based ensembles that encompass bagging techniques (Random Forest, ExtraTrees) 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 held-out 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

TABLE I: Aggregated Model Performance Metrics (Mean ± Std Dev across 5 Seeds)

Model	Test R^2 (Mean ± Std)	Test MSE (Mean ± Std)	Mean CV R^2 (Mean ± Std)	Mean CV MSE (Mean ± Std)
CatBoost	0.9075 ± 0.0162	378.5910 ± 101.9490	0.9100 ± 0.0100	501.1131 ± 40.1203
ExtraTrees	0.9028 ± 0.0266	423.1093 ± 67.9559	0.9205 ± 0.0099	433.0565 ± 61.9017
Random Forest	0.8946 ± 0.0342	419.5756 ± 143.8633	0.9149 ± 0.0119	459.0965 ± 62.6069
Gradient Boosting	0.8872 ± 0.0277	468.2575 ± 150.5223	0.9198 ± 0.0116	421.2588 ± 27.0704
XGBoost	0.8742 ± 0.0436	501.5635 ± 144.9626	0.9155 ± 0.0098	457.7594 ± 25.9370
AdaBoost	0.8737 ± 0.0570	513.4382 ± 160.9231	0.9195 ± 0.0141	431.8544 ± 46.8239
Linear Regression	0.8264 ± 0.0840	644.2401 ± 195.9131	0.8546 ± 0.0295	746.2765 ± 105.7173
Ridge	0.8256 ± 0.0697	675.4093 ± 218.7416	0.8625 ± 0.0237	719.1190 ± 80.4998
Lasso	0.8171 ± 0.0935	692.6660 ± 263.4297	0.8686 ± 0.0234	682.5610 ± 78.6915

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

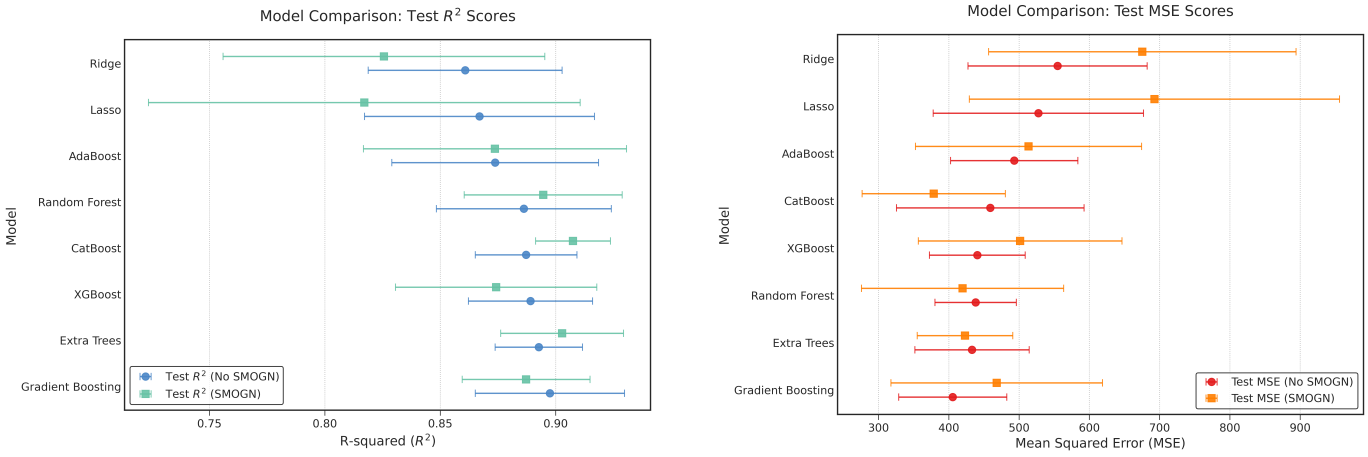
To rigorously evaluate the contribution of SMOGN data augmentation to model performance, an ablation study was conducted. This study mirrored the primary experimental setup in terms of data sources, preprocessing, the set of nine models, the 5-seed multi-run evaluation, and the use of Optuna for model 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.

IV. Results and Discussion

This section presents a comprehensive empirical evaluation of the proposed machine learning framework. First, it analyzes the performance of the nine regression models when trained with the joint SMOGN and hyperparameter optimization strategy. Following this, the section presents the findings of the ablation study, designed to quantify the specific impact of SMOGN augmentation. Finally, it explores the interpretability of the top-performing model and discusses the practical implications of the findings. All results are reported as the mean ± standard deviation across the five independent runs.

A. Performance of Models with Joint Optimization

The aggregated performance metrics for all nine models are summarized in Table I, with a visual comparison provided in Figure 2. The results clearly indicate that tree-based ensemble methods demonstrated strong and stable predictive capabilities when jointly optimized with SMOGN. CatBoost emerged as the top-performing model, achieving the highest mean Test R^2 of 0.9075 ± 0.0162, followed closely by ExtraTrees (0.9028 ± 0.0266) and Random Forest (0.8946 ± 0.0342). For the leading model, the error is also of practical importance. CatBoost's lowest mean Test MSE of 378.5910 ± 101.9490 translates to a Root Mean Squared Error (RMSE) of approximately 19.5 runs, offering a tangible error margin for player evaluation.



(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).

The strong and stable Test R^2 scores (generally above 0.89 for the top ensembles) confirm that the framework still achieves effective generalization to unseen data. In contrast, simpler linear models exhibited more modest performance, underscoring the necessity of using non-linear ensembles to capture the complex relationships within the augmented data. While a direct numerical benchmark against prior WPL research is challenging due to differing objectives, our framework's high predictive accuracy establishes a significant new benchmark for data-scarce sports analytics.

B. Ablation Study: Impact of the Framework

To rigorously isolate the impact of the joint optimization framework, an ablation study was conducted by tuning models on the original, non-augmented data. This experiment established a strong baseline, with Gradient Boosting emerging as the top performer, achieving a mean Test R^2 of 0.8975 ± 0.0323 and a Test MSE of 405.5981 ± 76.8926.

A direct comparison to this baseline confirms that the joint optimization framework yielded significant performance gains for key models. The most notable improvement was observed in CatBoost, whose Test R^2 increased by +2.3 percent (from approximately 0.8872 to 0.9075). This demonstrates that for certain architectures, the integrated approach of simultaneously tuning the model and the data augmentation process can deliver substantial predictive improvements.

However, the study also revealed a critical insight. The impact of the framework, while powerful, is highly model-dependent. For instance, the baseline top performer, Gradient Boosting, saw a decline in its R^2 with the introduction of SMOGN, while AdaBoost's performance remained nearly unchanged. This nuanced outcome highlights that the benefits of data augmentation are not absolute but are deeply intertwined with a model's learning algorithm. It underscores that the true power of the framework lies in its ability to discover these synergistic model-augmentation pairings through a joint, comparative approach.

C. Testing Generalizability: A Cross-Domain Case Study

To test the feasibility of our proposed pipeline, another dataset was used to train a subset of models, which responded best to SMOGN tuning on the WPL dataset. The models chosen were Gradient Boosting and Random Forest.

The dataset used is a medical insurance dataset [18] that has 6 features and 1,300 rows. Minor adjustments were made for the larger dataset. More aggressive pruning was used, and instead of generating synthetic data within the cross-validation loop for all training folds, SMOGN generated synthetic data once per trial. Additionally, only the relevance function was adjusted to accommodate the change in the data distribution, and the number of trials was reduced to 20. The generalizability of the framework was validated in the insurance data set, where the top models initially established a baseline test R^2 of ≈ 0.863 without SMOGN.

Applying the joint optimization framework, the performance changes depended on the model. For Gradient Boosting, the test RMSE increased from 4410.23 ± 2646.64 to 4696.32 ± 2513.58, corresponding to a 6.5% increase in error, while the test R^2 declined from 0.8637 ± 0.0310 to 0.8437 ± 0.0296, a 2.3% decrease in explained variance. In contrast, Random Forest exhibited improvements: its test RMSE decreased from 4428.33 ± 2668.28 to 4338.69 ± 1324.63, a 2.0% reduction, and test R^2 increased from 0.8626 ± 0.0318 to 0.8698 ± 0.0032 , an 0.8% improvement. Mirroring the WPL results, this effect depended on the model, reinforcing that the benefits of joint optimization are highly synergistic with specific model architectures. This cross-domain validation demonstrates the applicability of the framework beyond sports analytics.

D. Synthesis and Practical Implications

The results of the WPL case study and the cross-domain validation on the insurance dataset demonstrate that the proposed joint optimization framework is a generalizable approach for regression under data scarcity. Tree-based ensembles, including CatBoost and Random Forest, consistently outperformed linear and simpler non-linear models. These results confirm that the framework's effectiveness arises from jointly optimizing data augmentation and model hyperparameters rather than relying on domain-specific feature engineering. The synergy between augmentation and model tuning enables the identification of configurations that maximize predictive performance for each learning algorithm.

The framework also provides practical utility. In the WPL study, CatBoost achieved an RMSE of approximately 19.5 runs, giving actionable error margins for evaluating player performance and guiding team strategy. Cross-domain validation on the insurance dataset illustrates the framework's adaptability to different tabular datasets. The integrated approach further identifies which augmentation strategies best complement specific models, offering insights into how model performance interacts with data scarcity and synthetic data generation.

Overall, the framework delivers accurate and interpretable predictions while offering insights into performance drivers. By simultaneously optimizing data augmentation and model parameters, it provides an effective solution for regression in data-scarce domains such as sports analytics, finance, healthcare, and engineering. These results demonstrate that joint optimization can significantly enhance predictive modeling when data is limited.

V. Limitations

This study has a few inherent limitations. The primary constraint is the reliance on only three WPL seasons, which, despite augmentation, may limit generalizability and contribute to variability across evaluation runs. Additionally, although the 50-trial Optuna search for joint hyperparameter optimization was systematic, it may not have fully explored the vast parameter space required to reach the global optimum.

VI. Conclusion

This paper presents a machine learning framework for predicting season-long batting performance in the data-scarce

Women's Premier League. A key contribution is a joint optimization framework that integrates data augmentation with model hyperparameter tuning into a single system, representing a clear improvement over traditional sequential methods. Implemented using Optuna to concurrently tune SMOGN and nine regression models, multi-seed evaluations demonstrated that this integrated approach enables ensemble models, particularly CatBoost, to achieve high predictive accuracy and stability. The ablation study confirmed that the framework plays an essential role in achieving performance gains, although the extent of improvement varies across models.

This research makes two primary contributions. First, it presents a generalizable blueprint for enhancing AutoML pipelines by incorporating joint optimization of data synthesis and model selection, providing a general methodology for data-scarce regression applicable to diverse domains such as finance, healthcare, and engineering. Second, it validates the framework's practical utility by demonstrating its ability to provide actionable insights in data-scarce environments, using sports analytics as a high-impact testbed for tasks like player evaluation and performance management. Together, these contributions underscore the value of integrating data augmentation with model tuning, laying the groundwork for developing reliable and generalizable predictive systems in data-limited settings.

VII. Future Work

Future research can extend this framework in several directions. A primary avenue is to improve generalizability by validating the methodology on larger datasets, future WPL seasons, and other sports or data-limited domains, thereby testing the framework's adaptability as domain-specific dynamics evolve. Methodological enhancements could include incorporating additional contextual features, such as player form metrics, venue characteristics, or opposition analysis [19], which help capture domain-specific factors more effectively.

Another direction is to further enhance performance under data scarcity. This could involve exploring alternative data augmentation strategies beyond SMOGN, including generative methods such as Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs) for tabular data, and further optimizing SMOGN to reduce computational cost, enabling more extensive statistical testing. Integrating these augmentation techniques with Automated Machine Learning (AutoML) pipelines [20] could enable end-to-end optimization of feature selection, data augmentation, and model hyperparameters, improving both the efficiency and reliability of the framework.

Finally, while the current study focuses exclusively on regression, extending the framework to classification tasks in data-scarce settings remains an open challenge. Incorporating Explainable AI (XAI) methods [3] alongside such extensions would enhance interpretability and practical utility, providing deeper insights into performance drivers in sports analytics and other domains limited by data.

References

[1] 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

[2] 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*, 2019.

[3] 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)*, 2023, pp. 1–6.

[4] 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)*, 2023.

[5] 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*, 2023.

[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, 2024.

[7] Y. Agrawal and K. Kandhway, "Winner prediction in an ongoing one day international cricket match," *Journal of Sports Analytics*, vol. 9, pp. 305–318, 2024.

[8] 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)*, 2024, pp. 1–6.

[9] 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)*, 2023, pp. 344–348.

[10] C. D. Prakash and S. Verma, "A new in-form and role-based deep player performance index for player evaluation in t20 cricket," *Decision Analytics Journal*, vol. 2, no. 7, p. 100025, 2022.

[11] 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, 2023.

[12] 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, 2021.

[13] 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, 2022.

[14] 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, 2024.

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

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

[17] F. P. et al., "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] B. Lantz, *Machine Learning with R: Expert Techniques for Predictive Modeling*, 3rd ed. Packt Publishing, 2019.

[19] 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, 2024.

[20] 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, 2024.