

A Counterfactual Sensitivity-Based Learning Framework for Outcome Prediction in One-Day International Cricket

Rakesh V S¹, Bhavana P², Mamatha S³, Pushpanathan G⁴

¹Department of Computer Science and Engineering, ²Department of Computer Science and Engineering (IoT, Cyber Security including Blockchain Technology), ^{3,4}Department of Information Science and Engineering

^{1,2,3}Cambridge Institute of Technology, K R Puram, ⁴BMS Institute of Technology
^{1,2,3,4}Bengaluru, Karnataka, India

Abstract: Predicting outcomes in One-Day International (ODI) cricket is difficult due to rapidly changing match conditions and complex in-game dynamics. Traditional machine learning models often rely on static representations and fail to capture the sensitivity of outcomes to realistic variations in matches. This paper proposes a Scenario-Aware Counterfactual Sensitivity-Based Learning (SCIL) framework for predicting ODI match outcomes. The framework generates plausible counterfactual match scenarios and assigns influence weights based on prediction sensitivity, enabling the model to emphasize decisive match situations. Experiments on historical ball-by-ball ODI data show that SCIL consistently outperforms conventional models, including Logistic Regression, Random Forest, and XGBoost. The results show enhanced robustness, predictive accuracy, and probabilistic discrimination, highlighting the effectiveness of counterfactual sensitivity modeling for reliable cricket match prediction outcomes.

Keywords: Counterfactual Learning, Match Outcome Prediction, One-Day International Cricket, Scenario-Aware Modeling, Sports Analytics.

1. INTRODUCTION

Cricket, particularly the One Day International (ODI) format, generates abundant structured data that can be leveraged for predictive modelling and strategic decision-making. The variability inherent in match outcomes—driven by team composition, player form, environmental conditions, and in-match dynamics—presents a formidable challenge for traditional statistical approaches. Sophisticated machine learning (ML) methods have demonstrated substantial potential for addressing complex prediction tasks in sports analytics, yielding insights that inform performance optimization and tactical planning across disciplines. In recent work, dynamic predictive

frameworks have been proposed that model match progression across multiple game states and extract informative features via optimization techniques, yielding improved outcome-prediction accuracy relative to conventional approaches [1].

Simultaneously, advanced predictive modelling has been applied to individual performance components in ODIs, such as player scoring and bowler output, uncovering deeper interactions between contextual variables and performance metrics [2]. Studies in sports analytics increasingly integrate feature extraction, model optimization, and data-driven insights to refine predictions of match outcomes, thereby enabling stakeholders to make evidence-based decisions ranging from team selection to in-game strategy [3]. Moreover, robust predictive models that account for contextual attributes such as venue, toss outcome, and team strength are crucial for producing generalizable and interpretable results [4].

This work explores a novel application of machine learning to a large, ball-by-ball ODI cricket dataset to forecast match outcomes with greater reliability. By employing a carefully selected set of engineered features that reflect both pre-match and dynamic in-match conditions, the framework evaluates the power of predictive models tailored to the complexities of ODI cricket.

Main contributions of this study include:

- Development of a predictive framework using engineered features from ball-by-ball ODI cricket data for match outcome prediction.
- Incorporation of context-aware features, including toss decisions and team strength indicators, to enhance prediction accuracy and interpretability.
- Evaluation of multiple machine learning models on historical ODI data to analyze performance differences.

- Analysis of prediction outcomes to demonstrate practical relevance for match forecasting and strategic decision-making.

The remainder of the paper is organized as follows. Section 2 reviews recent literature on machine learning applications in cricket analytics and match outcome prediction. Section 3 explains the methodology adopted in this study, including the dataset description, feature construction, and the proposed learning framework. Section 4 presents the experimental results along with a detailed discussion of the findings. Section 5 concludes the paper by summarizing the major outcomes and suggesting directions for future work, followed by the references.

2. RELATED WORKS

Recent advances in cricket analytics reflect a clear transition from simple score-based summaries to detailed ball-by-ball modeling, where match context evolves rapidly and plays a crucial role in determining outcomes. With the growing availability of structured cricket datasets, machine learning techniques have been widely adopted to extract patterns in player performance, team behavior, and match results. These developments have laid a strong foundation for predictive and decision-support systems in cricket analytics.

Cricket analytics has evolved to focus on data-driven models that use extensive match records to link individual player metrics to team outcomes, supporting practical decision-making for performance insights [5]. In addition to traditional structured data, video analytics has gained importance, with deep learning techniques used to identify batting techniques from match footage, showing that cricket performance patterns can be effectively learned with appropriate representations [6]. From a temporal standpoint, probabilistic models have been used to analyze persistence and predictability in cricket, often referred to as the “hot hand” effect, underscoring the significance of time-dependent behavior in assessing performance [7].

Research efforts have also expanded to event understanding in broadcast content, where deep learning frameworks identify key match moments and generate player-specific highlights, emphasizing that match outcomes are only one part of a larger cricket intelligence ecosystem [8]. Recent studies have also examined pose estimation combined with machine learning to classify batting strokes using body key-points, thereby enhancing interpretability by connecting predictions to

biomechanical cues rather than relying solely on scoreboard data [9]. At the same time, context-aware decision-support systems have been proposed that incorporate match state and textual commentary to suggest bowling and fielding strategies, demonstrating how predictive models can assist with actionable in-game decisions beyond mere outcome prediction [10].

Dynamic outcome modeling has also received increasing attention, in which supervised learning techniques are applied to structured cricket-match features to forecast match outcomes and to compare the effectiveness of learning paradigms. Such studies demonstrate that outcome prediction accuracy can be significantly influenced by feature representation and learning strategy, particularly when match context is explicitly modeled [11]. Complementing ODI-focused studies, machine learning approaches have also been applied to shorter formats, such as T20 Internationals, demonstrating that both static and evolving match features contribute to predictive performance across cricket formats [12]. Additionally, several studies have focused on predicting specific match conditions, such as high-score chases, using machine learning, demonstrating that a combination of contextual match features and advanced classifiers can significantly improve predictive capability for dynamic in-game events [13].

Overall, existing studies indicate the growing importance of machine learning across various areas of cricket analytics, including performance assessment, strategy development, and outcome forecasting. However, many methods depend on static representations or analyze isolated parts of the game, which restricts their ability to account for real-world match situations. This gap underscores the need for scenario-aware learning frameworks that explicitly model how plausible changes in match conditions can affect predictions.

3. RESEARCH METHODOLOGY

This section describes the dataset preparation process, feature construction, and the proposed scenario-aware learning framework adopted for ODI cricket match outcome prediction. The methodology is designed to capture both historical performance patterns and contextual match dynamics, which are critical for reliable prediction in limited-overs cricket.

3.1. Dataset Description

The experiments are conducted using the One-Day International (ODI) cricket dataset from the Cricsheet

repository, accessed via the Kaggle platform [14]. The dataset provides structured ball-by-ball and match-level information for international ODI matches, enabling detailed analysis of match outcomes and performance patterns. Multiple comma-separated value (CSV) files are used to capture information at different levels, including match summaries, batting and bowling statistics, and player details, as shown in Table 1. Only completed matches with valid results are considered, while abandoned or incomplete matches are excluded to ensure data consistency and reliability.

Table -1: Dataset Files and Description

File name (.csv)	Level	Description
match_summary	Match	Match result, teams, venue, date, toss winner, toss decision
batter_player_stats	Player (Batting)	Runs scored, balls faced, strike rate, boundaries
bowler_player_stats	Player (Bowling)	Overs bowled, wickets taken, economy rate
detailed_player_data	Player	Player role and participation details
Ball-by-ball YAML (original Cricsheet)	Ball	Delivery-wise match events

The processed dataset comprises 2,400 completed ODI matches with clearly defined outcomes. In addition to match-level records, the dataset includes 1,540 aggregated batting performance entries, 929 aggregated bowling performance entries, and 52,031 player-level participation records. Table 2 presents statistics for the different CSV files. These multiple levels of granularity enable comprehensive modeling of both match context and player contributions.

Table -2: Dataset Statistics

File	Rows	Columns	Description
match_summary	2,400	21	Match-level details (teams, venue, toss, result)
batter_player_stats	1,540	16	Aggregated batting

			performance per match
bowler_player_stats	929	17	Aggregated bowling performance per match
detailed_player_data	52,031	21	Player participation and role information

External contextual factors such as weather conditions, pitch reports, and player injury information are excluded because these data are not consistently available across all matches in the dataset.

3.2. Data Preprocessing

Prior to model training, the raw data were preprocessed to ensure consistency and suitability for machine learning. Incomplete and abandoned matches were removed based on missing outcome labels. Player-level batting and bowling statistics were aggregated to the match level using standard summary statistics, including total runs, total wickets, average strike rate, and economy rate. Categorical variables, including team identifiers, venue, and toss decision, were encoded into numerical representations, while numerical features were normalized to reduce scale bias. Missing numerical values arising from partial player participation were imputed using feature-wise means, yielding a clean, unified match-level dataset.

3.3. Feature Construction

After preprocessing, a structured feature set was constructed to represent each ODI match as a fixed-length numerical vector suitable for supervised learning. Feature design focused on capturing match context, team performance, and progression patterns that influence outcomes in limited-overs cricket. All features were derived directly from the available dataset.

Match Context Features

Match context features denote static conditions linked to a game. These include team identifiers, venue details, and toss-related factors. Toss decision and winner are included as categorical variables because they affect batting and bowling strategies. Venue identifiers help indirectly capture location-specific effects.

Let c_k denote a categorical feature. Each such feature was mapped to a numerical representation via an encoding function, as shown in Eq. (1).

$$c_k \rightarrow \phi(c_k) \quad (1)$$

where $\phi(\cdot)$ denotes the encoding operation.

Batting Performance Features

Batting performance features summarize the overall effectiveness of the batting side at the match level. Player-level batting statistics were aggregated using standard summary operations.

For a given match m with N For batters, the total runs scored are computed using Eq. (2).

$$Runs_m = \sum_{i=1}^N r_i \quad (2)$$

where r_i is the number of runs scored by the batter i . Average batting strike rate is computed using Eq. (3).

$$AvgSR_m = \frac{1}{N} \sum_{i=1}^N sr_i \quad (3)$$

These indicators reflect scoring volume and scoring efficiency, respectively.

Bowling Performance Features

Bowling performance features capture the bowling side's ability to restrict runs and take wickets. Bowling statistics were aggregated at the match level.

Total wickets taken in a match are computed as per Eq. (4).

$$Wkts_m = \sum_{j=1}^K w_j \quad (4)$$

where w_j denotes wickets taken by the bowler j , and K is the number of bowlers. The average bowling economy is computed using Eq. (5).

$$AvgEco_m = \frac{1}{K} \sum_{j=1}^K eco_j \quad (5)$$

where eco_j is the economy rate of the bowler j in that match.

Match Progression Features

To model the dynamic nature of ODI matches, performance features were aggregated across key phases of an innings: powerplay, middle overs, and death overs. Using Eq. (6), the phase-wise run rates are computed.

$$RR_p = \frac{Runs_p}{Overs_p} \quad (6)$$

where p denotes the corresponding phase. $Runs_p$ and $Overs_p$ denotes the runs scored and overs bowled in the phase p . This representation allows the model to capture momentum shifts across different stages of the match.

All constructed features were combined into a fixed-length vector:

$$x_m = [x_1, x_2, x_3, \dots, x_d] \quad (6)$$

where x_m represents the feature vector for the match m , and d denotes the total number of engineered features. This vector serves as the input to the proposed learning framework described in the subsequent section.

3.4. Scenario-Aware Counterfactual Sensitivity Learning

The proposed method enhances match outcome prediction by incorporating realistic counterfactual scenarios during model training. A base predictive model is first trained using the original match instances. For each instance, multiple plausible counterfactual variants are generated through bounded and domain-consistent feature perturbations. The influence of each scenario is quantified by measuring the sensitivity of the model's predicted outcome to variations in that scenario.

Scenarios that produce significant shifts in predictions are assigned higher influence weights, whereas negligible variations are suppressed. Model parameters are then refined using a joint loss function that combines original instances with influence-weighted counterfactual scenarios. This learning strategy enables the model to capture outcome sensitivity to critical match dynamics and improves robustness to realistic match variations.

Equations (7)–(15) formally define the proposed counterfactual sensitivity-based learning mechanism, including scenario generation, influence estimation, and scenario-aware optimization.

Let the training dataset be

$$D = \{(x_i, y_i)\}_{i=1}^N \quad (7)$$

where $p_i = f_\theta(x_i)$ represents the match state and $y_i \in \{0, 1\}$ denotes the match outcome. A probabilistic prediction model $f_\theta \in \mathbb{R}^d \rightarrow [0, 1]$ is trained to estimate the outcome probability $p_i = f_\theta(x_i)$.

Counterfactual Scenario Generation

For each instance x_i , a set of S counterfactual scenarios are generated as

$$x'_{i,s} = x_i + \delta_{i,s}, \quad s = 1, 2, \dots, S \quad (8)$$

where the perturbation vector $\delta_{i,s}$ satisfies

$$\delta_{i,s}^{(j)} \in [-\dot{\phi}_j, +\dot{\phi}_j] \quad (9)$$

and all $x'_{i,s}$ are constrained to lie within the valid domain Ω .

Sensitivity-Based Influence Estimation

The predicted outcome for each counterfactual scenario is

$$p'_{i,s} = f_{\theta}(x'_{i,s}) \quad (10)$$

The sensitivity (prediction shift) induced by each scenario is defined as

$$\Delta_{i,s} = |p'_{i,s} - p_i| \quad (11)$$

A thresholded influence score is computed as

$$I_{i,s} = \max(0, \Delta_{i,s} - \tau) \quad (12)$$

where τ controls insignificant fluctuations.

Influence Weight Normalization

The normalized influence weight of each scenario is given by

$$w_{i,s} = \begin{cases} \frac{\exp(\lambda I_{i,s})}{\sum_{k=1}^S \exp(\lambda I_{i,k})}, & \sum_k I_{i,k} > 0 \\ \frac{1}{S} & \text{otherwise} \end{cases} \quad (13)$$

Scenario-Aware Learning Objective

The final training objective is formulated as

$$L(\theta) = \sum_{i=1}^N \left[\alpha \ell(f_{\theta}(x_i), y_i) + (1 - \alpha) \sum_{s=1}^S w_{i,s} \ell(f_{\theta}(x'_{i,s}), y_i) \right] \quad (14)$$

where $\ell(\cdot)$ denotes a standard classification loss function and $\alpha \in [0, 1]$ balances original and counterfactual learning. The optimal model parameters are obtained as

$$\theta^* = \arg \min_{\theta} L(\theta) \quad (15)$$

Collectively, these equations constitute the core of the proposed framework, ensuring that outcome predictions remain consistent and robust under realistic variations in counterfactual matches.

4. RESULTS AND DISCUSSIONS

This section evaluates the effectiveness of the proposed Scenario-Aware Counterfactual Influence Learning (SCIL) framework for outcome prediction in One-Day International (ODI) cricket. The performance of the proposed method is compared against conventional machine learning baselines to demonstrate the impact of counterfactual sensitivity modeling.

4.1. Experimental Setup

The experiments were conducted on historical One-Day International (ODI) cricket match data, where each sample represents an intermediate match state characterized by features such as overs remaining, runs required, wickets in hand, current run rate, and required run rate. The task is formulated as a binary classification problem to predict whether the batting or the chasing team ultimately wins the match.

The proposed Scenario-Aware Counterfactual Influence Learning (SCIL) framework is implemented using a probabilistic learning model and compared to widely used machine learning algorithms such as Logistic Regression, Random Forest, and XGBoost. All models are trained on the same feature set and follow the same experimental pipeline, with SCIL differing only in its use of counterfactual scenario generation and influence weighting based on sensitivity. Stratified train-test splitting is used to maintain class balance during evaluation. The models are developed and evaluated with standard machine learning libraries, applying consistent hyperparameter tuning strategies across all methods.

Performance is assessed using both classification and probabilistic metrics. Accuracy captures overall correctness, whereas Precision and Recall reflect the quality and completeness of positive-prediction outcomes. The F1-score provides a balanced measure of these two aspects, which is particularly important in closely contested match situations. In addition, the Area Under the Receiver Operating Characteristic Curve (AUC) is reported to evaluate the models' ability to rank winning and losing outcomes across varying decision thresholds, offering insight into the reliability of predicted probabilities.

4.2. Overall Performance Evaluation

The overall performance comparison emphasizes the ability of different models to capture decisive match dynamics beyond simple pattern learning. By jointly

assessing classification accuracy and probability ranking quality, the evaluation reveals the impact of counterfactual-sensitivity modeling. The comparison shown in Fig. 1 provides insight into how SCIL enhances the robustness of outcome prediction.

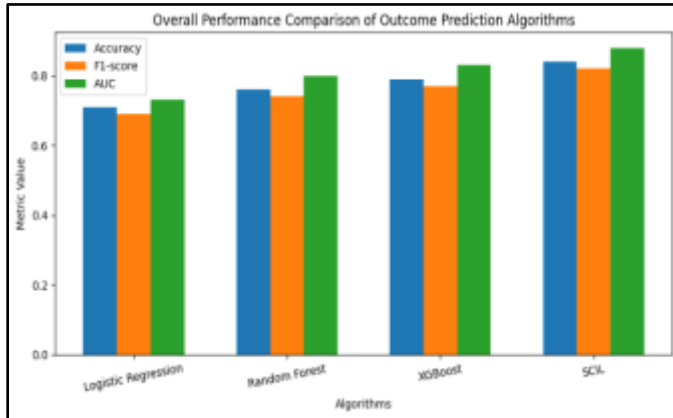


Fig. 1: Overall Performance Comparison of Outcome Prediction Algorithms

As illustrated in the figure, SCIL consistently achieves the highest performance across all evaluation metrics. While ensemble-based models such as Random Forest and XGBoost outperform Logistic Regression, the proposed SCIL framework further improves prediction accuracy and F1-score, indicating a better balance between precision and recall. SCIL achieves the highest performance across all metrics, attaining an accuracy of 0.84 and an AUC of 0.88, compared with 0.79 accuracy and 0.83 AUC for XGBoost, the strongest baseline model. The most notable improvement is observed in AUC, indicating that SCIL provides superior probabilistic discrimination between winning and losing outcomes. These results confirm that incorporating counterfactual sensitivity and influence-based weighting enables the model to capture critical match dynamics more effectively than conventional learning approaches, leading to more reliable outcome predictions in closely contested ODI matches.

4.3. Counterfactual Scenario Contribution Analysis

This analysis examines how individual counterfactual scenarios contribute to learning within the proposed SCIL framework. Since SCIL assigns influence weights based on prediction sensitivity, scenarios are expected to have unequal impact on outcome learning. Visualizing the distribution of these weights provides insight into how the

emphasis of learning is allocated across generated scenarios. Fig. 2 shows the analysis.

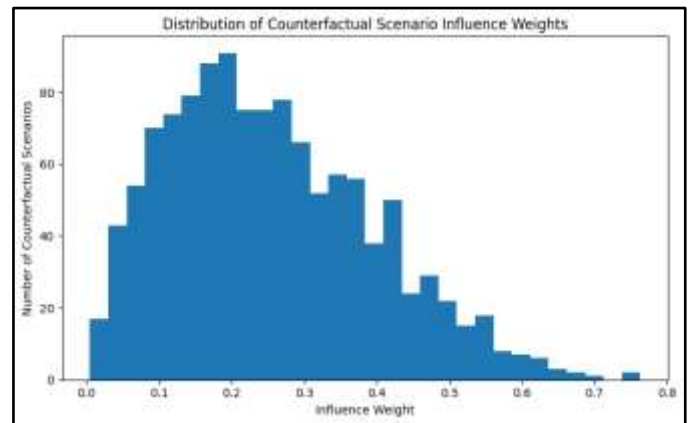


Fig. 2: Distribution of Counterfactual Scenario Influence Weights

The influence weight distribution exhibits a long-tailed pattern, in which only a small subset of counterfactual scenarios receives strong influence, whereas most contribute marginally. This indicates that match outcomes are driven by a limited number of decisive variations rather than uniform changes. By prioritizing these influential scenarios, SCIL effectively focuses learning on critical match dynamics, thereby improving robustness and predictive performance. A small fraction of the generated counterfactual scenarios receive high influence weights, whereas the majority contribute only marginally.

5. CONCLUSIONS

This study presented a Counterfactual Sensitivity-Based Learning (SCIL) framework for outcome prediction in One-Day International cricket, designed to capture the influence of realistic match variations on predictive behavior. By integrating sensitivity-weighted counterfactual scenarios into the learning process, the proposed approach improves robustness and interpretability compared to conventional models. Experimental evaluation shows that SCIL achieves superior performance, attaining an accuracy of 0.84 and an AUC of 0.88, outperforming strong baseline methods such as XGBoost, which achieved an accuracy of 0.79 and an AUC of 0.83. The analysis further demonstrates that only a small subset of counterfactual scenarios significantly contributes to learning, validating the importance of influence-based weighting. Future work will extend this framework to other cricket formats, incorporate player-level and contextual features, and

explore real-time decision-support applications for match strategy analysis.

REFERENCES

1. K. Dhinakaran and S. Anbuchelian, "Enhanced cricket match prediction using kernel methods for feature extraction and back-propagation neural networks," *Nature*, Jan. 2026, doi: 10.1038/s41598-026-36555-6.
2. S. Sanjaykumar, K. Udaichi, G. Rajendiran, M. Cretu, and Z. Kozina, "Cricket performance predictions: a comparative analysis of machine learning models for predicting cricket player's performance in the One Day International (ODI) world cup 2023," *Health Sport Rehabilitation*, vol. 10, no. 1, pp. 6–19, Mar. 2024, doi: 10.58962/hsr.2024.10.1.6-19.
3. Y. Agrawal and K. Kandhway, "Winner prediction in an ongoing one day international cricket match," *Journal of Sports Analytics*, vol. 9, no. 4, pp. 305–318, Jan. 2024, doi: 10.3233/jsa-220735.
4. R. Lokhande, R. N. Awale, and R. R. Ingle, "Forecasting bowler performance in One-Day International cricket using Machine learning," *Expert Systems With Applications*, vol. 259, p. 125178, Aug. 2024, doi: 10.1016/j.eswa.2024.125178.
5. M. J. Awan *et al.*, "Cricket match analytics using the big data approach," *Electronics*, vol. 10, no. 19, p. 2350, Sep. 2021, doi: 10.3390/electronics10192350.
6. T. Moodley, D. Van Der Haar, and H. Noorbhai, "Automated recognition of the cricket batting backlift technique in video footage using deep learning architectures," *Scientific Reports*, vol. 12, no. 1, p. 1895, Feb. 2022, doi: 10.1038/s41598-022-05966-6.
7. S. K. Ram, S. Nandan, and D. Sornette, "Significant hot hand effect in the game of cricket," *Scientific Reports*, vol. 12, no. 1, p. 11663, Jul. 2022, doi: 10.1038/s41598-022-14980-7.
8. R. Mahum, A. Irtaza, S. U. Rehman, T. Meraj, and H. T. Rauf, "A Player-Specific framework for cricket highlights generation using deep convolutional neural networks," *Electronics*, vol. 12, no. 1, p. 65, Dec. 2022, doi: 10.3390/electronics12010065.
9. H. U. R. Siddiqui *et al.*, "Enhancing Cricket Performance Analysis with Human Pose Estimation and Machine Learning," *Sensors*, vol. 23, no. 15, p. 6839, Aug. 2023, doi: 10.3390/s23156839.
10. A. Hussain, S. Arshad, and A. Hassan, "RunsGuard Framework: Context aware Cricket game strategy for field placement and score containment," *Applied Sciences*, vol. 14, no. 6, p. 2500, Mar. 2024, doi: 10.3390/app14062500.

11. P. Manikiran, N. Sri Ram, K.V.V. Abhilash, A. Vamsi Krishna, Prof. Mohammed Zabeeulla A N, "Cricket Match outcome prediction using Machine learning techniques", *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, vol. 11, Jun. 2022, doi: 10.17148/IJARCCE.2022.11619.
12. P. Sudasinghe, S. Viswakula, and P. Lakraj, "Can social media opinions add value to historical data?: A study for T20I cricket match outcome prediction using machine learning," *Journal of Sports Analytics*, vol. 11, May 2025, doi: 10.1177/22150218251342185.
13. S. A. A. Shah, Q. Zaman, D. Wasim, J. Allohibi, A. A. Alharbi, and M. Shabbir, "Optimal model for predicting highest runs chase outcomes in T-20 international cricket using modern classification algorithms," *Alexandria Engineering Journal*, vol. 114, pp. 588–598, Dec. 2024, doi: 10.1016/j.aej.2024.11.113.
14. Decentralized, "Cricsheet ODI cricket data," Kaggle, <https://www.kaggle.com/datasets/decentralized/cricsheet-odi-cricket-data>.

BIOGRAPHIES



Rakesh V S is an Assistant Professor in the Computer Science & Engineering Department at Cambridge Institute of Technology, K R Puram, Bengaluru, with ten years of teaching experience. He earned both his B.E. and M.Tech degrees from Visvesvaraya Technological University and is a certified KSET scholar. His research has been published in international journals and conferences, earning several best paper awards. His areas of interest encompass machine learning, data mining, network security, compiler design, and algorithms.



Bhavana P is an Associate Professor in the Department of Computer Science & Engineering (IoT, Cybersecurity including Blockchain Technology) at Cambridge Institute of Technology, K R Puram, Bengaluru, with 15 years of teaching experience. She holds a B.E. in Information Science and Engineering and an M.Tech. in Computer Science and Engineering from Visvesvaraya Technological University, Belagavi. Her areas of expertise include data analytics, database management systems, computer architecture, system software, and compiler design. She is a Life Member of ISTE.



Mamatha S is an Assistant Professor in the Department of Information Science & Engineering at the Cambridge Institute of Technology with more than 9 years of academic experience. Her areas of expertise include object-oriented programming, data structures, web technologies, and Python programming. She holds an M. Tech in Computer Networks from Visvesvaraya Technological University and a B.E. from the Cambridge Institute of Technology. She is currently pursuing her Ph.D. in interdisciplinary areas of AI and healthcare.



Pushpanathan G is an Assistant Professor in the Department of Information Science & Engineering at BMS Institute of Technology, Bengaluru, with over 9 years of experience and academic interests in data structures and algorithms, operating systems, big data, and automata theory. His expertise also includes UNIX shell programming, web technologies, and system modeling and simulation. He holds a B.E. and an M.Tech. in Computer Science from Visvesvaraya Technological University. He has actively participated in numerous workshops and faculty development programs.