JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# IPL Match Win Predictor

**[1]Pratyush Kala, [2]Amita Goel, [3]Vasudha Bahl, [4]Nidhi Sengar**
[1]Student, [2]Head of Department, [3]Assistant Professor, [4]Assistant Professor
[1]Department of Information Technology,
[1]Maharaja Agrasen Institute of Technology, Delhi, India

***Abstract :*** This study applies machine learning (ML) and deep learning (DL) techniques to predict Indian Premier League (IPL) match outcomes using historical data. Traditional models often struggle with the complexity of cricket due to its many influencing factors. To improve prediction accuracy, a hybrid classification approach is used, incorporating Random Forest (RF), Gradient Boosting (GB), Support Vector Classifier (SVC), and Deep Neural Network (DNN). Data from IPL matches (2008–2024) is preprocessed through encoding, feature engineering, and scaling. Models are trained and tuned using Scikit-learn and TensorFlow. Results show that RF and GB achieved the highest accuracy (around 86.78%), while the DNN showed potential for further refinement. The findings highlight the value of ensemble and DL methods in modeling match outcomes and support their use in real-time prediction frameworks

***IndexTerms*** **- ML, DL, IPL Prediction, Ensemble Models, DNN, SVC, RF, GB, Sports Analytics, Feature Engineering, Cricket Data, Match Outcome Forecasting, Scikit-learn, TensorFlow.**

---

## 1.INTRODUCTION

The Indian Premier League (IPL) has emerged as one of the most widely followed cricket tournaments, drawing global attention and a large fan base. Predicting the outcomes of IPL matches is particularly challenging due to the many variables at play—ranging from team strategies and player form to toss results, venue dynamics, and weather conditions. Conventional methods such as linear regression and basic decision trees often struggle to handle the complex; non-linear relationships present in this domain.

One of the core difficulties in making accurate IPL predictions is accounting for the ever-changing nature of the game, including shifts in team composition, seasonal variations, and evolving playing conditions. This research proposes a comprehensive prediction system that leverages both machine learning and deep learning approaches, utilizing historical match data to improve forecast reliability.

The main goals of this study include designing a deep learning-based model to estimate match outcomes, analyzing critical predictors like recent team form, venue performance, toss results, and head-to-head statistics, and benchmarking the model's performance against standard baseline classifiers such as logistic regression and decision trees.

The paper is organized as follows: Section 2 reviews existing literature, Section 3 describes the methodology and dataset, Section 4 presents the results and discusses key findings, and Section 5 concludes with suggestions for future research.

## 2. Literature Review

### 2.1 Existing work:

Kapadia et al. [1] conducted a comparative study of various machine learning algorithms for predicting IPL match results. Their work emphasized the effectiveness of classifiers like Random Forest, Naive Bayes, Model Trees, and K-Nearest Neighbors (KNN), with feature selection handled via filter-based methods. Evaluation using metrics such as precision, recall, and accuracy demonstrated that tree-based models performed more reliably than statistical and probabilistic counterparts. Notably, they found the inclusion of toss outcomes introduced inconsistencies, and they proposed further exploration into how machine learning can advance sports analytics.

In another study, Kampakis and Thomas [2] utilized historical data from the English Twenty20 Cup to develop predictive models based on an extensive set of over 500 team and player metrics. Their analysis involved diverse classification techniques and feature selection strategies. Results indicated that gradient-boosted decision trees outperformed other models, reinforcing the dominance of tree-based methods over more traditional statistical approaches.

Mahajan et al. [3] examined IPL match prediction through supervised learning by incorporating features such as team form, home-ground advantage, and individual player statistics. Algorithms including Random Forest, Naive Bayes, KNN, and Gradient Boosted

Decision Trees were applied to assess team and player performance. Their findings underscored the precision of ensemble models and provided practical insights for improving prediction strategies in cricket analytics.

Bandulasiri [4] investigated outcomes of One-Day International (ODI) cricket matches using logistic regression. Factors like home-field benefit, match type, toss decisions, and fielding conditions were analyzed. The research also considered the Duckworth-Lewis method in rain-affected scenarios and assessed its reliability. The study offered a deeper understanding of how contextual game variables influence match results.

Passi and Pandey [5] aimed to predict individual player performances, particularly batting and bowling outcomes, in ODI cricket. Using several classifiers—including Naive Bayes, Random Forest, Support Vector Machines (SVM), and Decision Trees—they concluded that Random Forest consistently yielded the most accurate predictions across both batting and bowling metrics.

Ahmed [6] employed a multivariate data mining approach to predict ODI match outcomes, focusing on Pakistan's national team. Variables like team rankings, toss results, venue and weather conditions, and recent performance streaks were considered. A wide array of machine learning techniques, including ANN, Logistic Regression, Random Forest, and KNN, were evaluated for their classification performance

Sinha [7] built models to predict IPL match results by analyzing game-specific features such as venue, teams involved, and toss outcomes. Their study experimented with six machine learning techniques—Decision Tree, Naive Bayes, Random Forest, ANN, Logistic Regression, and KNN—and reported Random Forest as the most accurate (88.46%). They also innovatively incorporated sentiment analysis of Twitter data to reflect public perception of teams and players.

Finally, Ahmed et al. [8] focused on the performance variability of Pakistan's cricket team in ODIs, using features like strike rate, batting/bowling averages, economy rates, and fielding metrics. Their model comparisons showed SVM achieving the highest accuracy (82.5%), with batting average and strike rate emerging as the strongest predictors of match success.

### 2.2 Identified Gaps:

Despite the progress made in cricket match prediction models, several gaps persist in the current body of research:

- **Static Feature Limitation:** Many studies focused primarily on static variables such as team composition and historical match statistics, overlooking recent team form, head-to-head performance, and venue-specific trends.
- **Limited Exploration of Toss Impact:** Although toss outcomes were included in some models, their true impact on match outcomes was not adequately explored, often leading to model inaccuracies.
- **Underutilization of Deep Learning Models:** Despite their strength in modeling complex, non-linear relationships, deep learning models like artificial neural networks (ANN) have not been widely used in this domain.

### 2.3 Contribution to the Field

This study addresses these issues by presenting a broader predictive framework that incorporates deep learning and expands the range of variables considered. Key innovations of this research include:

- **Incorporation of Recent Team Form:** Capturing team momentum and performance trends over the last five matches to account for current form.
- **Head-to-Head Win Percentages:** Analyzing historical matchups between teams to identify patterns and psychological advantages.
- **Venue-Specific Win Percentages:** Factoring in the impact of venue conditions and team familiarity with specific grounds.
- **Impact of Toss Outcome:** Investigates how toss outcomes influence match decisions, aiming to detect and correct hidden biases.
- **Average Runs in Previous Matches:** Assessing team batting performance and consistency over recent games to inform match predictions.

By integrating these variables into a deep learning framework, this research advances cricket match prediction methodologies and provides a more holistic approach to analyzing match outcomes.

## 3.METHODOLOGY

The study utilized historical IPL data covering the years 2008 to 2023, acquired from Kaggle. This dataset comprised both match-level summaries and detailed ball-by-ball information, offering a strong basis for in-depth analysis.

During the feature engineering process, key variables such as team momentum, toss outcomes, venue influence, and recent performance statistics were extracted to improve model accuracy.

Next, the data was pre-processed by encoding categorical variables and normalizing numerical ones. The dataset was then split into two parts: 80% for training the models and 20% for testing their performance to ensure fair evaluation. To build the predictive models, four classification techniques were selected: Random Forest, Gradient Boosting, Support Vector Classifier (SVC), and Deep Neural Network (DNN), each chosen for their strengths in handling classification tasks. All models were fine-tuned through hyperparameter optimization—using GridSearchCV for tree-based models, kernel adjustments for SVC, and dropout layers with the Adam optimizer for DNN.

Finally, model effectiveness was gauged using evaluation criteria such as accuracy, precision, recall, and F1-score. This approach enabled a meaningful comparison to determine which method was most suitable for forecasting IPL match outcomes.


***Fig 1. Methodology Flowchart***

# 4. Results & Discussions

## 4.1 Results:

| Model | Accuracy (%) | Remarks |
|---|---|---|
| Gradient Boosting Classifier | **86.78** | Highest accuracy; best at capturing patterns |
| Ensemble Voting Classifier | 82.84 | Combines strengths of multiple models |
| Random Forest | 76.62 | Strong performance; handles non-linear features |
| Deep Neural Network (DNN) | 67.66 | Moderate accuracy; requires more tuning/data |
| Support Vector Classifier | 54.73 | Weakest performance; limited by linear kernel |

*Fig 2. Model Results Table*

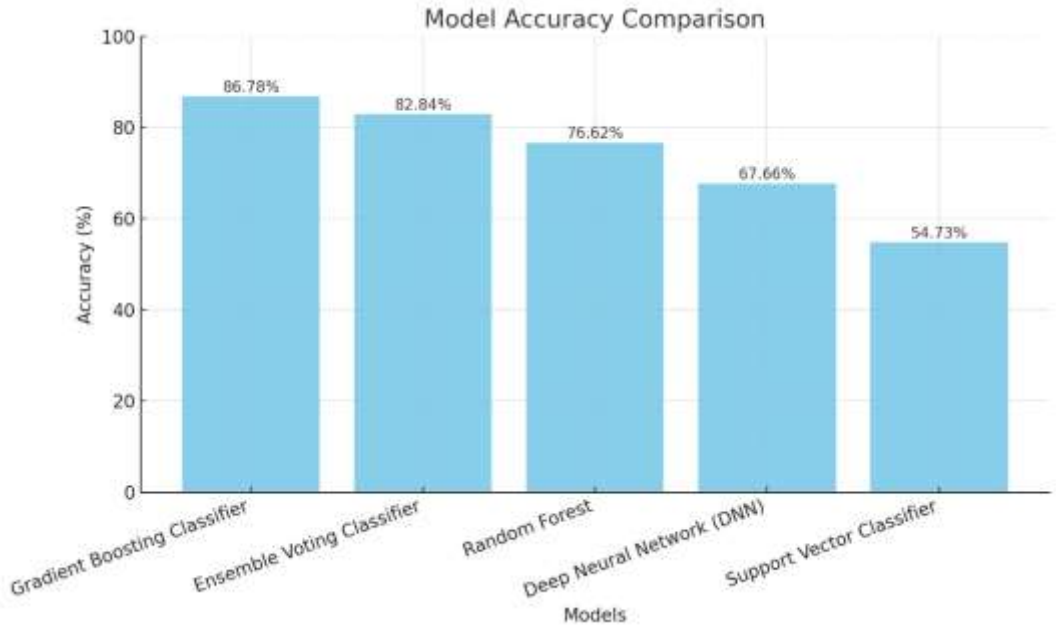


*Fig 3. Model Accuracy Comparison*

The comparative analysis of model performance shows that **tree-based ensemble methods** are the most effective for IPL match prediction. The **Gradient Boosting Classifier** stood out with the highest accuracy of 86.78%, likely because of its strong capability to capture intricate feature patterns and reduce the risk of overfitting.

The **Ensemble Voting Classifier**, which combines predictions from multiple models, followed closely with **82.84%** accuracy, suggesting that hybrid approaches can further enhance predictive reliability. The **Random Forest** model also performed well, indicating that decision tree ensembles are particularly well-suited to the structure and nature of the cricket match data.

In contrast, the **Deep Neural Network (DNN)**, though capable of capturing non-linear relationships, achieved only **67.66%** accuracy. This suggests the need for either a more extensive dataset or deeper architecture optimization to fully leverage the potential of deep learning. The **Support Vector Classifier (SVC)** showed the weakest performance at **54.73%**, indicating that linear models struggle with the high dimensionality and variability inherent in sports data.

In summary, the results emphasize how combining multiple models can improve performance in sports outcome prediction and underline the importance of fine-tuning algorithms and selecting relevant features.

## 4.2 Limitations:

**Real-Time Updates:** The model does not account for real- time player injuries, lineup changes, or unexpected match-day factors.

**Data Bias:** Historical data may introduce bias due to the dominance of certain teams or players in specific IPL seasons
**Toss Dependency:** While toss impact is included, its effect varies depending on match circumstances, which the model may not fully capture

# 5. CONCLUSION

.

This study developed a predictive framework leveraging machine learning techniques to forecast Indian Premier League (IPL) match outcomes, based on data spanning from 2008 to 2023. It utilized multiple ensemble models, including Gradient Boosting and Random Forest, in addition to Deep Neural Networks and Support Vector Classifiers. Model effectiveness was assessed using evaluation criteria such as accuracy, precision, recall, and F1-score. The results indicated that ensemble approaches generally performed better than other methods, with Gradient Boosting showing particularly strong results. The findings underscore the potential of data-driven approaches in sports analytics, particularly when applied to structured and time-evolving datasets. Future work will explore real-time prediction, model interpretability, and the integration of live match data to enhance forecasting accuracy.

**Future Scope**

**Incorporate Real-Time Player Statistics:** Integrate player form, fitness, and injury status to enhance prediction accuracy.

**Include Weather and Pitch Conditions:** Incorporate external factors such as weather and pitch conditions, which often influence match outcomes.

**Advanced Ensemble Models:** Explore hybrid models that combine **DNNs, Gradient Boosting, and Random Forests** to optimize prediction accuracy further.

## REFERENCES

**[1]** Kapadia K, Abdel-Jaber H, Thabtah F, Hadi W. Sport analytics for cricket game results using machine learning: An experimental study. Applied Computing and Informatics. 2020 Jul 28;18(3/4):256-66.

**[2]** Kampakis S, Thomas W. Using machine learning to predict the outcome of English county twenty over cricket matches. arXiv preprint arXiv:1511.05837. 2015 Nov 18.

**[3]** Mahajan MS, Kandhari MG, Shaikh MS, Pawar MR, Vora MJ, Deshpande MA. Cricket Analytics and Predictor.

**[4]** Bandulasiri A. Predicting the winner in one day international cricket. Journal of Mathematical Sciences & Mathematics Education. 2008;3(1):6-17.

**[5]** Passi K, Pandey N. Increased prediction accuracy in the game of cricket using machine learning. arXiv preprint arXiv:1804.04226. 2018 Apr 9.

**[6]** Ahmed W. A multivariate data mining approach to predict match outcome in one-day international cricket. M.S. Dissertation, Karachi Institute of Economics and Technology, Pakistan. 2015 Aug.

**[7]** Sinha A. Application of Machine Learning in Cricket and Predictive Analytics of IPL 2020.

**[8]** Ahmed W, Amjad M, Junejo K, Mahmood T, Khan A. Is the performance of a cricket team really unpredictable? a case study on Pakistan team using machine learning. Indian Journal of Science and Technology. 2020 Sep 24;13(34):3586-99.