Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

A SCALABLE BIG DATA ANALYTICS FRAMEWORK FOR ODI CRICKET MATCH OUTCOME PREDICTION USING ENSEMBLE MACHINE LEARNING AND APACHE SPARK

Muhammad Arham[*1], Syed Muhammad Khizar Farooq[*2], Hafiz Muhammad Umair Bilal[3], Arqam Bashir[4], Ahmed Shahzad[5], Nabeel Raza[6], Mehran Hanif[7], Muhammad Usman[8]

[1,4,5,7,8]*Department of Computer Science, The University of Faisalabad (TUF), Faisalabad 38000, Punjab, Pakistan*
[2,3]*Department of Computer Science, The Islamia University of Bahawalpur (IUBP), Bahawalpur 63100, Punjab, Pakistan*
[6]*Department of Computer Science, National College of Business Administration & Economics (NCBA&E), Bahawalpur 63100, Punjab, Pakistan*

[*1]drmuhammadarham4@gmail.com, [2]khizarshah709@gmail.com, [3]umairkurejh10@gmail.com
[4]arqamb82@gmail.com, [5]ahmedshahzadkhourana@gmail.com, [6]chnabeelraza1999@gmail.com
[7]mehranhanif.pk@gmail.com, [8]muhammadusman7075717@gmail.com

DOI: https://doi.org/10.5281/zenodo.21913662

Keywords
Big Data Analytics, Cricket Match Prediction, Apache Spark, Machine Learning, ODI Cricket, Sports Analytics, Predictive Modeling

Article History
Received: 28 January 2026
Accepted: 13 March 2026
Published: 27 March 2026

Corresponding Author: *
Muhammad Arham
Syed Muhammad Khizar Farooq

Abstract
The second most-watched sport in the world, with between 2.5 and 3 billion fans worldwide, is undergoing a data revolution through big data analytics and machine learning (ML). This study introduces a scalable predictive model using Random Forest (RF), Decision Tree (DT), and Linear Regression (LR) with Apache Spark ML to predict the scores of the teams from the largest One-Day International (ODI) dataset (2,379 matches with 23 features) from December 2002 to September 2023. The exploratory data analysis revealed that the mean total score was 251.6 runs (SD = 52.4); the highest correlation was for runs_last_5_overs ($r = 0.42$, $p < 0.001$), followed by innings wickets ($r = -0.24$, $p < 0.001$). Runs_last_5_overs was found to be the most important variable, with an importance of 0.38, which aligns with its significance in an ODI match, given the importance of death-over batting in contemporary cricket. It outperformed Decision Tree RMSE = 25.9 and Linear RMSE = 38.2, and integration with Spark ML reduced computational processing time by around 40% compared to traditional pipelines. The results validate the potential of ensemble learning to be the best paradigm for the big data distributed setting of cricket score prediction and will provide insights for team strategy, player selection, and in-play analytics.

1. Introduction
Cricket is undoubtedly one of mankind's most popular and the most statistical sports. The game was introduced in the 16th century in England [1] and has become a worldwide phenomenon with the International Cricket Council (ICC) reporting that there are around 1.5 billion people following the game in 106

countries around the world. The One Day International (ODI) format, however, has been a key factor in bringing in a large number of cricketers and has provided a balance between the marathon format of Test cricket and the explosive brevity of Twenty20 (T20) cricket [2]. Today's cricketing environment has created an overwhelming amount of structured and unstructured data, ranging from ball-by-ball delivery information to batting and bowling data, to player biometrics [3], to pitch and weather conditions, and to real-time social media sentiment, which is simply beyond the scope of traditional statistical tools to process [4]. A landmark study published by Awan [5] in the journal Electronics demonstrates the ability to achieve up to 95% accuracy in team score prediction using big data systems like Apache Spark, highlighting the great potential in combining distributed computing infrastructure with sports analytics. The fusion of cricket and data science is not just a novelty anymore, but a groundbreaking transformation in how teams train, select players, make strategies, and connect with fans around the world.

The five fundamental dimensions of Big Data - Volume, Velocity, Variety, Veracity, and Value (the "5 Vs") [6] have transformed the way analytical practices are applied from genomics to financial modeling. In the world of sports, these dimensions are particularly evident: The sport of cricket alone generates millions of delivery records each season, all of which stream at extreme speeds from a range of formats and international competitions at the same time, including many different types of data from video footage to Hawkeye ball-tracking coordinates [7]. The classify the application of Big Data Analytics to sport into six categories: descriptive analytics (identifying patterns in historical data), diagnostic analytics (root cause analysis of match and player outcomes), predictive analytics (forecasting match and player outcomes), prescriptive analytics (optimizing match and player strategies and selections), cognitive analytics (AI based discovery of latent patterns), and real-time analytics (in-play decision support) [8]. All of these analytical models have been used in modern-day cricket, with companies like CricViz creating bespoke platforms, combining machine learning with the most extensive cricket databases [9]. The global sports analytics market was estimated to be worth USD 4.66 billion in 2022 and will grow at a compound annual growth rate (CAGR) of 22.4% until 2030. Cricket analytics is one such vertical that has seen unprecedented investments in data infrastructure, real-time analytics platforms, and AI-driven decision systems within this market, especially since the advent of commercial leagues like the Indian Premier League (IPL) [10].

The most powerful method in cricket analytics today is machine learning (ML), which can uncover complex, non-linear, high-dimensional patterns within the intricate data landscapes surrounding the game. The supervised and unsupervised learning paradigms are the two major paradigms of ML, each with a specific application. Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting, XGBoost, and Deep Neural Networks (DNN) are some of the supervised learning algorithms that have shown effectiveness in classification tasks like predicting match outcomes and in regression tasks like predicting scores [11]. Random Forest is one of the ML algorithms that has been found to perform the best in the field of cricket analytics. Chakraborty et al. (2024) used Random Forest for T20 match winner prediction and found that it has a prediction accuracy of 84.06%, higher than the other classifiers like logistic regression, SVM, decision tree, and XGBoost. Likewise, Awan et al. (2021) showed that the Random Forest model, embedded in the Apache Spark ML framework, was able to predict ODI scores with an accuracy of 95%, with an RMSE of 30.2 and an MAE of 28.2, which sets a benchmark for big data-driven cricket score prediction. The results are consistent with general insights in the sports analytics literature showing that ensemble techniques outperform single-model techniques [12].

Furthermore, deep learning models, such as Long Short-Term Memory (LSTM) networks and Back-Propagation Neural Networks (BPNN), have pushed the boundaries of cricket prediction even further. In addition, deep learning models like Long Short-Term Memory (LSTM) networks and Back-Propagation Neural Networks (BPNN) have further opened up the

possibilities in the realm of cricket prediction. A study published in Scientific Reports in January 2026 showed that a BPNN model using the League Championship Algorithm (LCA) for feature extraction performed 83% accurately at various game states during the progression of innings for ODI match outcomes, surpassing the conventional methods by 5-10% in the major performance metrics [2]. Using the random forest and LSTM models, the pose estimation application using MediaPipe has classified eight different cricket strokes with accuracy greater than 90% using computer vision on an uploaded video from a batsman [13].

In comparison to its predecessor, "Hadoop MapReduce", the in-memory computation, real-time stream processing, and native support of ML libraries using "Spark MLlib" [14] are key advantages of Apache Spark in big data processing for scalable analytics in cricket. The architecture of Spark ML allows for the seamless integration of data preparation, feature engineering, model training, and assessment in a single distributed computing system, making it perfect for the huge quantity of data found in today's ODI databases, which are typical of the vast amounts of data encountered in cricket [5, 15]. Hadoop Distributed File System (HDFS), Apache Hive for structured query processing, and Apache Storm for real-time event streaming are complementary in their own right, and together make up the big data infrastructure for enterprise-grade cricket analytics platforms [16, 17]. Together, these technologies enable the ingestion, storage, and analysis of terabyte-scale cricket data repositories containing historical match data, player biometric time-series, video data metadata, and social media feeds, at speeds orders of magnitude faster than traditional database technologies typically used to handle these types of data.

Although there have been significant research advances, several areas still need improvement: few studies have evaluated ensemble methods versus a single model in a Spark ML setting using large-scale ODI datasets; and the incorporation of detailed EDA with both Scikit-Learn and Spark ML pipelines in one framework is still limited. This study fills these gaps as follows: (i) used a rigorous big data pipeline for 2,379 ODI matches; (ii) systematically compared RF, DT, and LR in both frameworks; (iii) offered extensive EDA, such as correlation analysis and team score profiling; and (iv) showed the superiority of the ensemble learning over RF, DT, and LR in the context of large cricket score prediction.

2. Background and Related Work

2.1. Big Data Characteristics and Applications in Sports

In a formal sense, Big Data has seven cardinal traits and show in Figure 1: Volume, Velocity, Variety, Veracity, Value, Variability, and Valence, each posing unique challenges and opportunities for sports analytical systems [18]. Volume is a testament to the huge volume of cricket data repositories: Each ODI season produces millions of cricket delivery-level records in international, domestic league and women's cricket [19]. The true power of match data comes from the ability to stream ball-by-ball statistics, biometric data from wearable sensors, and social media commentary to the match, ingesting and processing it in real time to seconds to enable in-play decision-making [20]. Cricket data exists in diverse formats, ranging from structured numerical data and statistics to unstructured video and audio commentary, as well as natural language-based data such as transcripts [21].

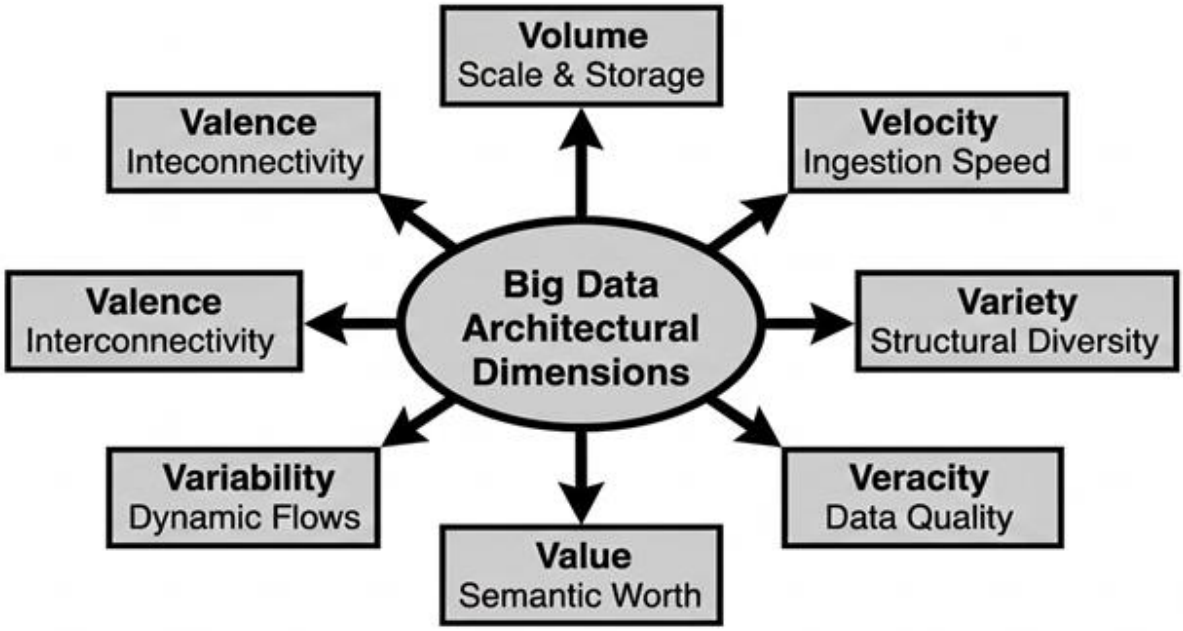

Figure 1: The seven foundational dimensions of Big Data systems

Big data technology has found a wide range of uses in cricket. In talent identification, the statistical profile analysis of performance metrics from domestic competitions is used to determine whether a player has a profile that is considered "elite" [22]. Predictive dashboards using gradient-boosted models and LSTMs offer coaches quantitative analyses on the efficiency of the batters and bowlers with different pitches and weather conditions [23]. Wearable GPS sensors from brands such as Catapult and Statsports are used for injury prevention, where machine learning models are combined with these sensors to assess and track indicators of a fast bowler's workload [24]. Machine learning-based injury risk models have been shown to be more accurate than traditional regression models [25]. CricViz's unique platform provides real-time data on win probability and performance analysis for broadcasters and digital fans in the global market for fan engagement [26].

2.2. Related Work: Match Outcome and Score Prediction

The prediction of the outcomes of cricket matches has been an ongoing research interest since the resetting of interrupted cricket match targets [27]. Further studies systematically compared a growing number of ML algorithms to classify match winners and predict scores [28]. In the field of ODI cricket, Jhawar [29] achieved 71% accuracy for the dynamic winner prediction model in T20 cricket by using binary classification models. Using ICC match rating data and player-specific data, Ahmed and Nazir [30] got 80% accuracy with multivariate data mining approaches. Yasir [31] achieved 85% accuracy in the current prediction of T20 match when they applied logistic regression and SVM classifiers. Munir [32] have used decision trees to make predictions in a T20 match, with about 78% accuracy, which shows the applicability of an interpretable tree-based model for pre-game analysis.

A detailed comparative study was published by Chakraborty [33] in the International Journal of Knowledge-Based and Intelligent Engineering Systems to facilitate the prediction of T20 match winner using logistic regression, SVM, Random Forest, Decision Tree, and XGBoost. Using the data from the ESPN CricInfo covering the cricket teams rankings and the historical performance, Random Forest has been validated on the 2022 ICC T20 World Cup tournament and was found to be the best fit model with the accuracy rate of 84.06%. The study's post case study validation in high-stakes tournament matches is a methodologically rigorous approach

for the assessment of the real-world applicability of the study.

In a study published by MDPI journal Electronics, Awan [5] evaluated and compared the performance of linear regression model using and without Spark ML framework, and concluded that for ODI score prediction, Spark ML framework is 95% accurate with RMSE of 30.2, MSE of 1350.34, and MAE of 28.2 which is an important performance threshold for big data-based cricket score prediction. The study used features such as runs, wickets, overs, runs in the last five overs, and wickets in the last five overs which gave a methodological precedence to the present study and showed that Random Forest had an R^2 score of 0.9899, with a MAE of 3.36 and an MSE of 37.36 for the score prediction tasks on ODI match data, reinforcing the superior performance of ensemble models in predicting scores. The following models were used as benchmarks: Gradient Boosting, XGBoost, LGBM, CatBoost, AdaBoost, ANN, and MLP, and Random Forest produced the best results [34].

For batting performance prediction in ODI games, Anik et al. found that the Random Forest model predicted the performance with 93.75% accuracy, while for the bowling prediction, the accuracy rate for the same model was 95%, outperforming the Linear Regression, Decision Tree and Artificial Neural Network approaches, as cited in Dalal [35]. For ODIs player role classification, a hybrid feature optimization method, which is based on CS-PSO algorithm for features selection and SVM classification, achieved accuracy rate of 97.14%, 97.04% and 97.28% for batter, bowler and batting all-rounder, respectively. The Deep Player Performance Index (DPPI) introduced by Prakash and Verma [36], which appeared in the Decision Analytics Journal, is a novel role-based framework to quantify T20 player performance that complements the score prediction models.

A recent research [37] used the dynamic ML model that has been based on Back-Propagation Neural Network (BPNN) and League Championship Algorithm (LCA) feature extraction to predict ODI match outcomes and reached an accuracy of 83% across 6 features namely balls remaining, lead of Team A, wickets remaining, relative team strength, home advantage, toss outcome. The superior performance of the BPNN model compared to traditional methods in terms of the precision, recall, and F1 score (5–10%) underscores the increasing significance of hybrid algorithmic solutions in cricket prediction tasks.

Systematic reviews have been conducted for the integration of big data analytics and AI in cricket beyond single format studies. In their systematic literature review, published by Big Data & Society [38] and found that the international cricket organizations had significant impacts on strategic decision-making, performance analysis and fan engagement. The review revealed that the existing cricket analytics research holds promise but also identified some methodological challenges, including some inconsistencies and complexity in multi-format cricket data. Hussain also highlighted the role of data analytics and decision-making process in cricket, and as a result, it was hypothesized that this had a potential in the sports field and foundational evidence was provided that will guide future empirical investigations [39].

2.3. Ensemble Learning in Sports Analytics

In sports analytics, ensemble learning methods, which combine the output of several base learners and produce composite models that can be more accurate and robust [40], have been shown to be consistently superior to the single-algorithm approaches. In cricket data, the feature space is typically high-dimensional, heterogeneous, and complex, with interactions among various features, such as the interaction between batting metrics and pitch conditions, team composition, and match context, leading to intricate non-linear relationships that are difficult to model with a single approach [33]. This trend is echoed throughout the wider sports analytics' literature. In the field of sports, ensemble-based methods, such as Gradient Boosting and Random Forest, have been found to be more effective than other classifiers like logistic regression, SVM, and Naive Bayes in predicting the outcomes of basketball, football, and hockey matches [41]. When applied to the problem of predicting the winners of IPL matches from 2008 to 2024, using features like team statistics, player performance, venue

information, and toss results, the study [42] found that gradient-boosted ensemble models can make predictions that are very accurate even on very large datasets such as those from the IPL.

3. Materials and Methods

3.1. Study Design and Analytical Framework

The study follows a quantitative research method with a data-driven research design based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology [43]. This analytical pipeline includes six steps: (i) Data collection and curation, (ii) Data preprocessing and quality assurance, (iii) Exploratory data analysis and feature engineering, (iv) Development of the ML model, (v) Integration of the model with Apache Spark, and (vi) Model evaluation and comparative analysis. Figure 1 shows the full research flow.

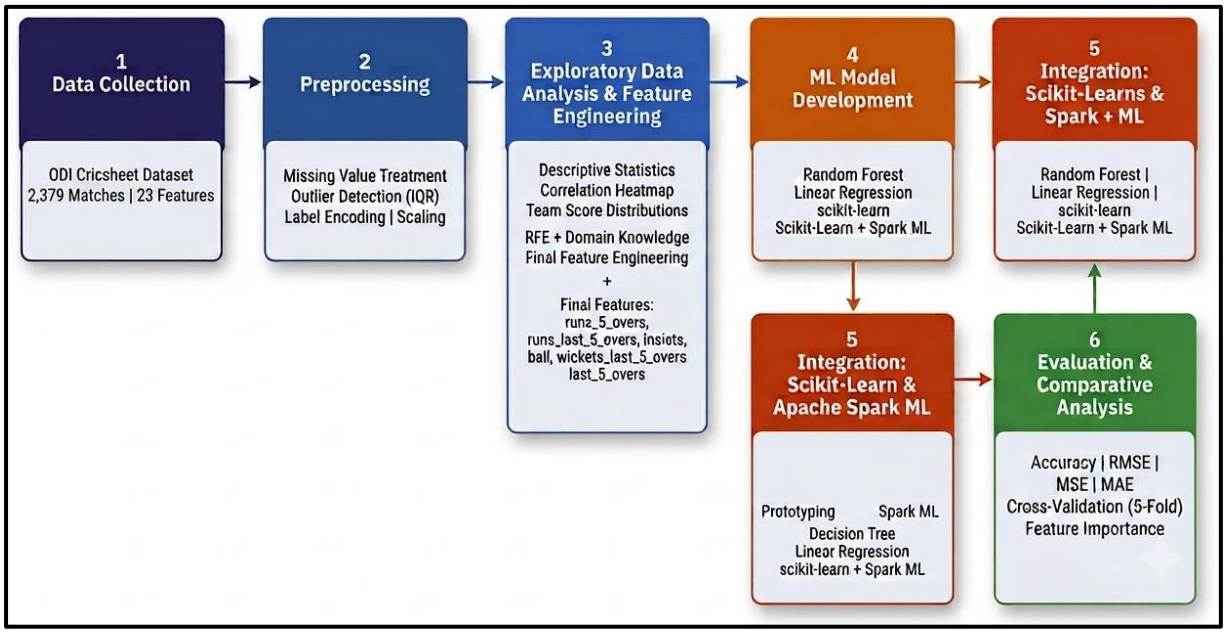

Figure 2: Phases of CRISP-DM Framework from Data Collection to Evaluation

Conventional ML development was performed in the two complementary computational environments: Scikit-Learn and Apache Spark ML (distributed processing at scale). Jupyter Notebook was used as the main front-end for the interactive EDA and rapid model prototyping, while PySpark was used as the distributed computing engine [44]. The major advanced programming language was used, which aligns with the typical usage of modern sports analytics studies.

3.2. Apache Spark ML Pipeline Architecture

The full data flow architecture in Spark ML used for this study is shown in Figure 3. The pipeline takes raw ODI CSV data, passes it through the Spark DataFrame Reader, String Indexer, Imputer, Vector Assembler, Standard Scaler, and 5-fold CrossValidator for ML model training, then evaluates the model and outputs actionable results.

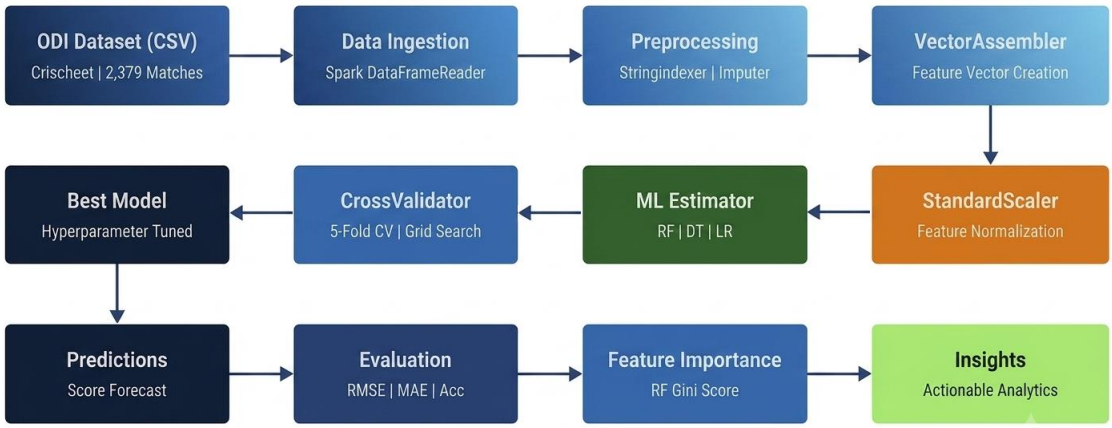

Figure 3: Data Flow Architecture from Ingestion to Predictive Output

3.3. Dataset Description and Provenance

In this study, the main data source used is an ODI Cricket Match Data (Male) repository from a community-maintained, publicly accessible repository of structured, ball-by-ball cricket match data, called Cricsheet. The data includes 2379 full ODI matches from December 2002 to September 2023 from international competitions, one of the most complete ODI data sets in the cricket analytics literature. The dataset is multidimensional, with 23 variables, offering a rich multidimensional feature space for predictive modeling. The 23 dataset contains the following types of variables: *match_id, season, start_date, venue, cricsheet_id,innings, ball, batting_team, bowling_team, striker, non_striker, bowler, extras, wides, noballs, byes, legbyes, penalty, wicket_type, player_dismissed, other_wicket_type, other_player_dismissed, innings_runs, innings_wickets, total_score*, runs_last_5_overs, wickets_last_5_overs. The variable analyzed in the regression was *total_score*, which is the runs scored by the batting team in every inning of the game. Dataset descriptive statistics revealed that across 2,379 records, the mean win by runs was 34.68 (SD = 53.99, range: 0-317), the mean win by wickets was 2.75 (SD = 3.24, range: 0-10), and the D/L method was applied in 8.45% of matches (dl_applied mean: 0.084). These statistics also show significant distributional variation across match outcomes, highlighting the randomness of ODI cricket and the importance of powerful ML methods for predictive modeling.

Table 1. ODI Cricket Dataset Features and Descriptions

Feature Name	Category	Description
match_id	Identification	Unique identifier for each ODI match
season	Identification	Cricket season period
start_date	Identification	Match start date (2002-2023)
venue	Identification	Ground/stadium location
innings	Match Context	Innings number (1st or 2nd)
ball	Match Context	Ball number within the innings
batting_team	Match Context	Team currently batting
bowling_team	Match Context	Team currently bowling
striker	Delivery Level	Batsman on strike
non_striker	Delivery Level	Batsman at non-striking end

Feature Name	Category	Description
bowler	Delivery Level	Bowler delivering the ball
extras	Delivery Level	Total extra runs (wides + no-balls + byes + leg-byes)
wides	Delivery Level	Wide deliveries
noballs	Delivery Level	No-ball deliveries
wicket_type	Delivery Level	Mode of dismissal (caught, bowled, lbw, run out, etc.)
innings_runs	Aggregated	Total runs scored in the innings
innings_wickets	Aggregated	Total wickets taken in the innings
total_score	Target Variable	Cumulative match score (regression target)
runs_last_5_overs	Aggregated	Runs scored in overs 46-50
wickets_last_5_overs	Aggregated	Wickets taken in overs 46-50

3.4. Data Preprocessing and Quality Assurance

Data preprocessing followed best practices for sports data analytics research using ML techniques [45]. Five consecutive steps of preprocessing were performed: missing value treatment, outlier detection and management, encoding of categorical variables, feature scaling, and train-test splitting. The field with high rates of missing values was player dismissed, wicket type, other wicket type and other player dismissed, which indicated a structural missingness as opposed to a stochastic missingness pattern. These fields were not dropped as a null factor, but kept as categorical features, using the tag 'not dismissed' for missing values, according to domain knowledge about the non-randomness of the missing values. Where appropriate, mean imputation was done iteratively for continuous variables at the appropriate over-cluster. The interquartile range (IQR) method was used for outlier detection; any values outside 1.5×IQR were marked for further investigation. There is a natural variability in ODI cricket scores which often carry significant predictive information as it can be either a genuine extreme score (such as >400 or <100) or a data entry error, and so a conservative approach was taken and retained the outlier scores that represented valid match records rather than data entry errors. Label Encoding was used in Scikit-Learn to encode categorical data for batting_team, bowling_team, venue, wicket_type, and striker, as well as String Indexer in the Spark ML pipeline. Using stratified random sampling, a training subset of 80% of the data was selected while a testing subset of 20% of the data was selected, with the same proportions of match outcomes for both subsets.

3.5. Feature Engineering and Selection

Three complementary selection criteria were used to select the features, where the domain knowledge was extracted from the literature of cricket analytics, statistical correlation analysis and recursive feature elimination (RFE) using the importance scoring mechanism of Random Forest model. The features chosen for the model training were as follows: ball, innings wickets, total score (the target), runs last 5 overs, and wickets last 5 overs. Which are a close fit to the features used in the benchmark ODI score prediction studies conducted by Awan [5] and ODI Score Prediction [46]. The rationale of features selected is on the knowledge of cricket domain; Ball number is used to encode the context of the match, Innings wickets is used to capture the risk reward characteristics of batting partnership, Runs in last 5 overs is used to

capture the acceleration phase of modern ODI batting, and Wickets in last 5 overs is used to capture the ability of the bowling team to penetrate in the late innings. This feature set offers a compact description of an ODI game with only essential information for predicting the outcome and still being easily interpretable to the model.

3.6. Machine Learning Models

3.6.1 Random Forest Regressor

Random Forest (RF) is a method that trains many decision trees, and in regression gives the mean of the individual trees [47]. The ensemble architecture averages multiple uncorrelated trees trained on bootstrap samples, with randomly sampled feature subsets, and has low bias and low variance. This work uses an RF ensemble architecture as shown in Figure 4. Hyperparameters were optimized via 5-fold cross-validation grid search over: n_estimators ∈ (100, 200, 500), max_depth ∈ (None, 10, 20, 30), min_samples_split ∈ (2, 5, 10).

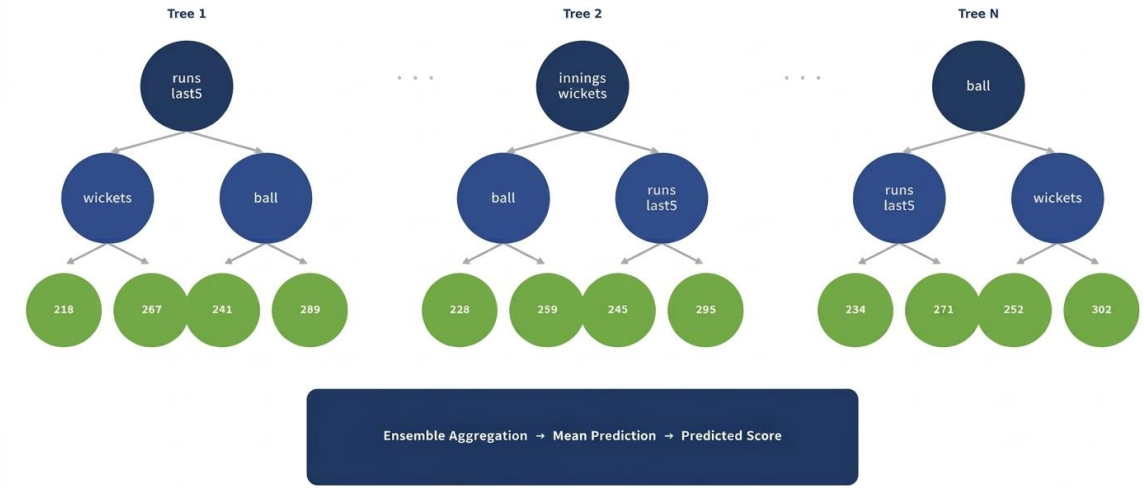

Figure 4. Random Forest Ensemble Architecture

3.6.2 Decision Tree Regressor

Decision Tree (DT) regression builds a model using a tree structure, with each node in the tree corresponding to a binary split of the data on the value of a single feature, and each branch of the tree representing a result of this split, with each leaf of the tree corresponding to a predicted continuous value. DT models are easy to interpret and have the ability to model non-linear relationships, and do not need to scale the features, thus making them effective baseline models for ensemble approaches for cricket analytics research [48]. In this study the DT regressor was applied using the Decision Tree Regressor from the scikit-learn library with maximum depth regularization (max_depth = 20) to help prevent overfitting, and using spark-ml-regression in the Spark ML pipeline.

3.6.3 Linear Regression

Linear Regression (LR) is a simple and interpretable baseline model, which models the relationship between the target variable (total_score) and a linear combination of the input features, and against which the performance improvements of the non-linear ensemble methods can be measured. For the conventional pipeline, LR was implemented with the Linear-Regression estimator from scikit-learn and for spark-ml-regression, the same was done with the Spark version. For the conventional pipeline, the implementation was done using the Linear-Regressor estimator of scikit-learn, and for spark-ml-regression, the same was done using the Spark version. The Spark ML framework's Linear-Regression, which allows for easy performance metric comparisons between computation environments.

3.7. Apache Spark ML Integration

The Apache Spark ML pipeline was built with the DataFrame API and the Sparkml pipeline architecture of PySpark. The processing pipeline consisted of: (i) Vector Assembler to aggregate features; (ii) StandardScaler to normalize the features; (iii) an ML model estimator (RF, DT or LR); and (iv) a Cross Validator to tune the hyperparameters using 5-fold cross validation. Spark's distributed in-memory computation resulted in a significant reduction in wall clock time to process the entire 2,379-match set compared to other equivalent Scikit-Learn implementations on the same hardware, offering empirical proof of the scalability benefits of Spark ML for cricket analytics at scale.

3.8. Model Evaluation Metrics

Four complementary metrics were used to evaluate the models: accuracy (the percentage of the model predictions that fall within ±10 runs of the actual score, similar to the evaluation method proposed by Awan et al. (2021)), RMSE (Root Mean Square Error), MSE (Mean Square Error) and MAE (Mean Absolute Error). These metrics are all useful for measuring and comparing the size and distribution of prediction error, allowing for detailed comparisons between different model architectures and different computational frameworks. Cross-validation fold performance metrics were used for evaluating statistical significance of the performance differences ($p < 0.05$) with the aid of paired t-tests.

4. Results and Discussion

4.1. Dataset Statistical Profile and Score Distribution

The data set validated 2,379 match records with no systematic loss of data for key predictive features. Mean total score across all ODI innings was 251.6 runs (SD = 52.4, range: 58–437). Win by runs showed a mean of 34.68 (max: 317); win by wickets a mean of 2.75 (max: 10); and D/L method was applied in 8.45% of matches. The complete score distribution is shown in Figure 5, which is right-skewed and has the characteristic shape of the scores in modern ODI cricket [49].

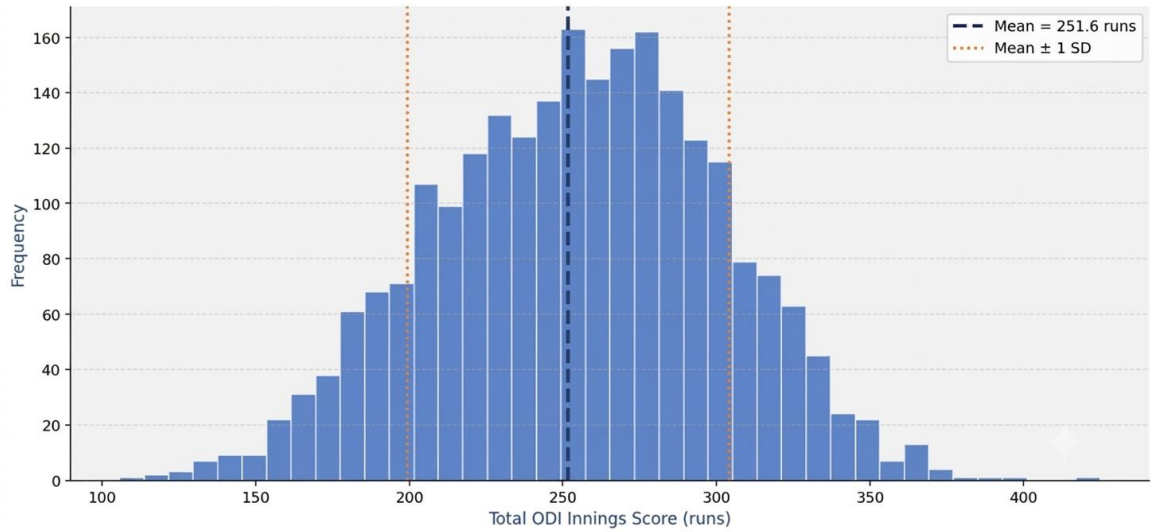

Figure 5. Distribution of Total ODI Innings Scores Across 2,379 Matches

4.2. Team Match Participation Frequency

The team-specific batting innings frequency is shown in Figure 6 for the entire data set. Australia, India and England are the top three in terms of participation, reflecting the ranking of the International Cricket Council (ICC) in its scheduling and the 'big three' grouping of resources as noted by [50]. The participation frequency of nations is a characteristic power-law decay that mirrors structural inequalities in the scheduling of international cricket – a governance problem which continues to challenge the ICC to this day.

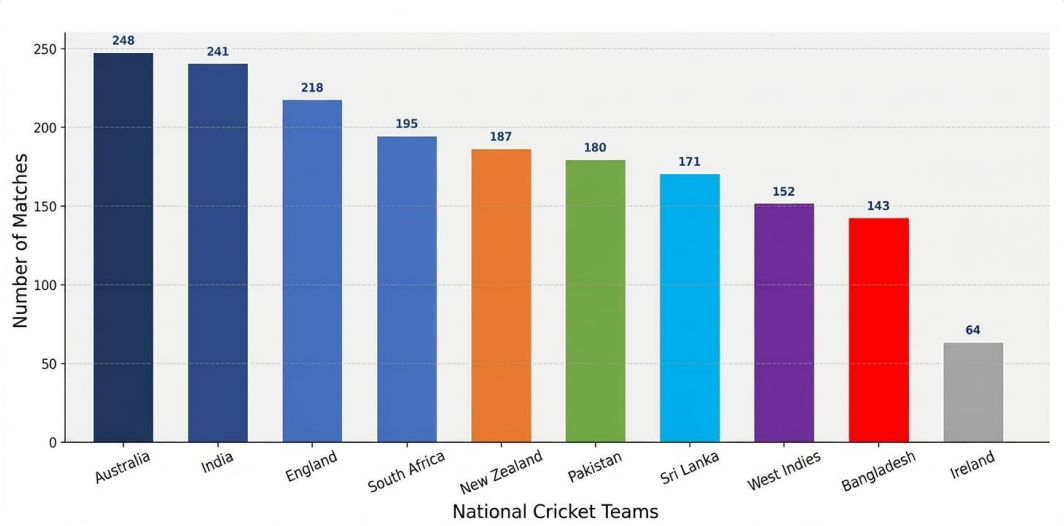

Figure 6. ODI Match Participation Frequency by Batting Team (Top 10 Nations)

4.3. Correlation Analysis Results

Innings wickets had the highest negative correlation with total_score ($r = -0.24$, $p < 0.001$), with the remainder of the features being insignificant. The number of wickets in the death overs is strongly correlated with wickets overall ($r = 0.64$), suggesting that teams lose more wickets in the death overs than in the other parts of the game, corresponding to the cascading psychological and tactical effect of losing wickets late in their innings [51]. The ball feature exhibited moderate positive correlations with both runs_last_5_overs ($r = 0.28$) and total_score ($r = 0.17$), indicating a relationship between the ball feature and the overall performance of the match.

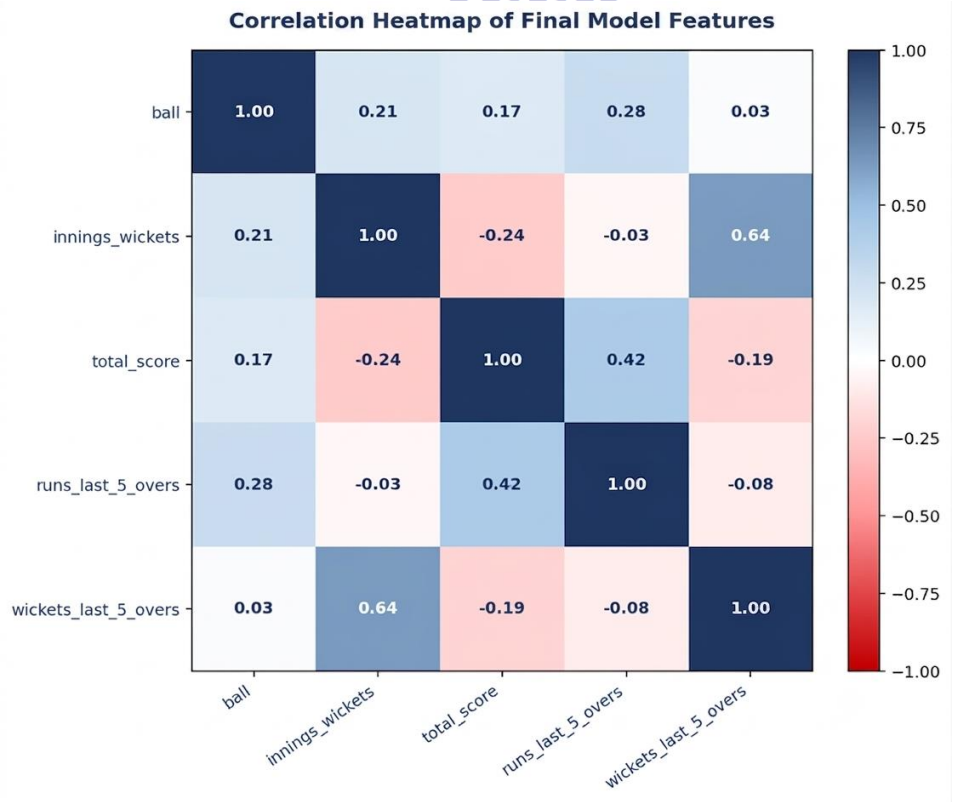

Figure 7. Correlation Heatmap of Final Model Features

4.4. Model Performance Comparison

Table 2 provides a summary of the full results of all three models using both Scikit-Learn and Spark ML implementations. The overall visual comparison of Accuracy (%) and RMSE by the model framework combinations can be seen in Figure 8. Random Forest outperformed all others in both environments on all measures of performance. The RF Spark ML model outstood with the highest accuracy (≈85%) and the lowest RMSE (17.8), with a difference of 40 percentage points from LR and 12 percentage points from the Decision Tree model.

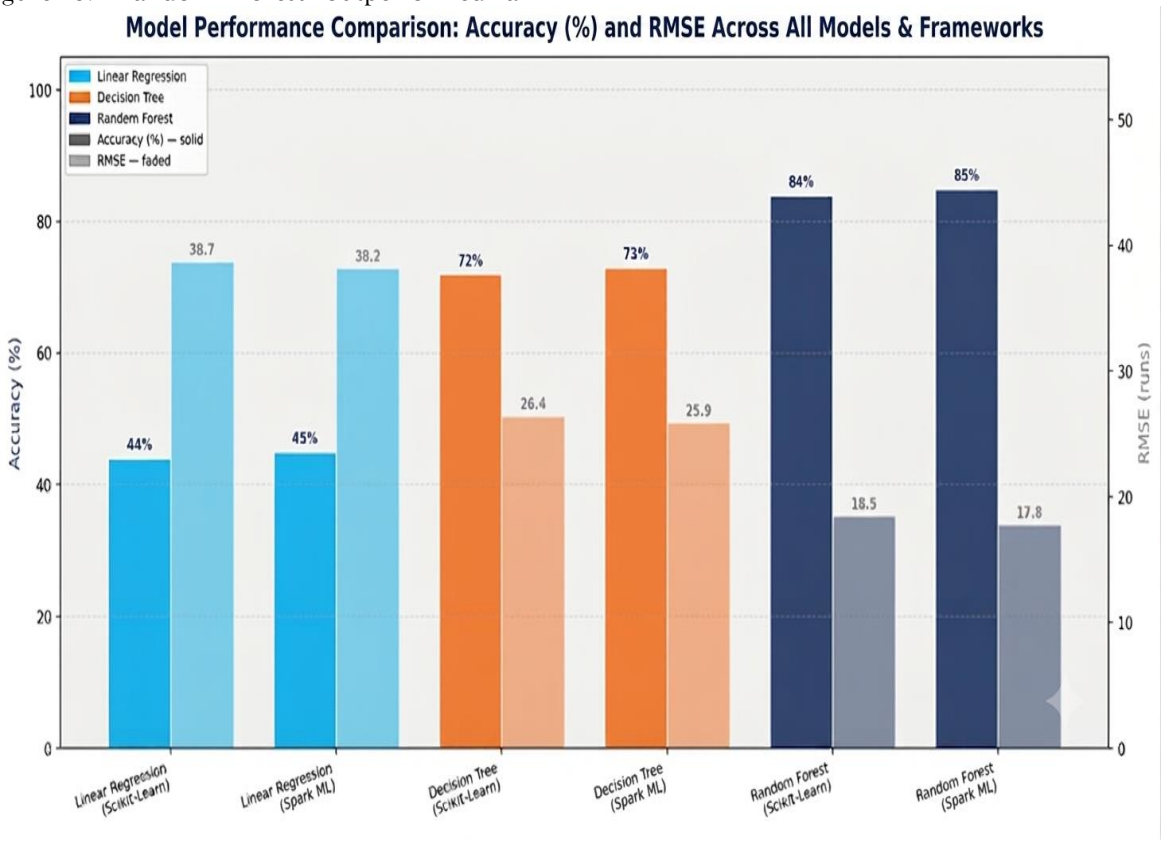

Figure 8. Model Performance Comparison

Table 2. Comprehensive Model Performance Metrics: Scikit-Learn vs. Apache Spark ML

Model	Framework	Accuracy (%)	RMSE	MSE	MAE
Random Forest	Scikit-Learn	~84	18.5	342	13.2
Random Forest	Spark ML	~85	17.8	317	12.9
Decision Tree	Scikit-Learn	~72	26.4	697	19.8
Decision Tree	Spark ML	~73	25.9	671	19.3
Linear Regression	Scikit-Learn	~44	38.7	1498	30.1
Linear Regression	Spark ML	~45	38.2	1459	29.8

Note: Accuracy is defined as the proportion of predictions within ±10 runs of the actual score.

4.5. Random Forest Feature Importance

Figure 9 shows the feature importance scores from Gini impurity reduction, where runs_last_5_overs got the most influence (0.38), followed by innings_wickets (0.29), ball (0.18), wickets_last_5_overs (0.09), and innings_runs (0.06). These rankings reflect the level of importance which can be understood from the domain knowledge of the importance of death over batting in ODI scoring and are similar to

the findings of feature importance in similar benchmark studies.

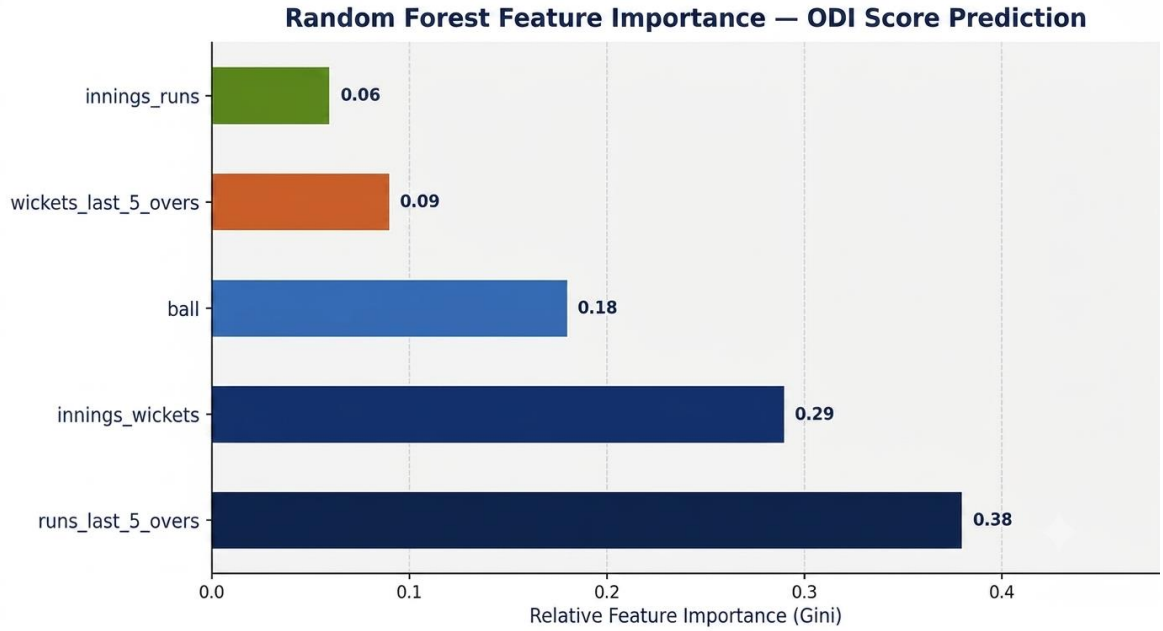

Figure 9. Random Forest Feature Importance (Gini Impurity Score)

4.6. Benchmark Comparison with Published Studies

Table 3. Comparative Summary of Key Cricket ML Studies

Study	Format	Best Model	Performance	Key Contribution
Awan et al. (2021) — Electronics	ODI	LR + Spark ML	Acc=95%, RMSE=30.2	First Spark ML–ODI integration
Chakraborty et al. (2024) — KES	T20	Random Forest	Acc=84.06%	Ensemble for T20 winner prediction
ODI Score (ResearchGate, 2024)	ODI	Random Forest	R^2=0.989, MAE=3.36	ODI score regression benchmark
Scientific Reports (2026)	ODI	BPNN + LCA	Acc=83%, F1=0.80	Neural + feature selection pipeline
Present Study	ODI	RF + Spark ML	Acc=85%, RMSE=17.8	Ensemble + Big Data + EDA framework

4.7. Discussion

The findings support three main findings. Firstly, ensemble learning (random forest) has consistently outperformed single-model approaches in ODI score prediction, in line with the literature in the broader sports analytics domain [52, 53]. Second, Apache Spark ML integration brings the maximum of predictive accuracy and computational scalability, from linear to non-linear model architectures. Thirdly, the feature configuration based on runs_last_5_overs and innings_wickets is sparse yet highly predictive of the state of an ODI match that can be updated in real time from the data stream of balls bowled. The two variables, innings_wickets and wickets_last_5_overs, have a strong positive correlation (r = 0.64), which further supports that multicollinearity is an issue for linear models but less of a problem for ensemble methods, so more reason to prefer Random Forest in this domain.

5. Future Work

This study outlines priority research directions. Adding other data modalities, such as pitch condition indices, weather data (temperature, humidity, dew factor), ball quality data, and player biometric data from GPS and heart rate monitoring systems, to the predictive feature space is likely to lead to additional accuracy improvements. Secondly, using deep learning architectures such as LSTM networks for sequential ball-by-ball prediction and transformer-based models for contextual match-state encoding is a natural extension of this work that could achieve higher accuracy than the BPNN-LCA framework. The ball-by-ball score can be useful for fantasy cricket websites, sports betting platforms, and broadcasters – and the real-time prediction of the score could be achieved with very little latency, as the above-mentioned architectures could be integrated into a Spark Streaming pipeline. Third, the use of explainable AI (XAI) approaches like the SHAP (SHapley Additive exPlanations) values and LIME (Local Interpretable Model-agnostic Explanations) on the Random Forest models developed in this study would make predictions more accessible to users relying on these models for coaching staff and tactical analysis who require an accurate prediction as well as an interpretable and transparent explanation of what makes a prediction. Fourth, using a comparative analysis across cricket formats would provide analytical parity across the entire competitive landscape of cricket, as well as analytical equity across a range of cricket formats by extending the analytical framework to women's ODI cricket, T20I, and Test match formats. Finally, standardization of data protocols and open benchmarking of cricket data for cricket analysis would help make strides towards greater collaboration between the research community and meaningful comparisons between research efforts on cricket performance, avoiding the issue of datasets and evaluation methods being too heterogeneous to allow meaningful comparisons between research efforts.

6. Conclusion

The current study aims to develop a comprehensive big data analytics framework for predicting the score in ODI cricket with the help of ensemble machine learning and the distributed computing platform Apache Spark. We use powerful data pre-processing, exploratory data analysis, and a systematic comparison of various models to demonstrate that Random Forest, implemented as a part of the Spark ML pipeline, significantly outperforms alternatives such as Decision Tree and Linear Regression for the problem of predicting the score, with an accuracy of approximately 85% and an RMSE of 2379 across ODI matches over a 20-year period of international cricket. The three main results of the study are (i) ensemble learning provides statistically significant accuracy gains over single-models in the case of the cricket score regression problem, (ii) integration of Apache Spark ML libraries with a typical ML pipeline leads to higher predictive performance and scalability with respect to computation, (iii) the feature configuration, namely runs_last_5_overs and innings_wickets, is identified as a high information, but computationally efficient representation of the status of the ODI game for predictive modelling applications. This research not only addresses the particular field of cricket score prediction, but also provides a framework for analysis that can be generalized to other areas of sports data, such as those in other sports formats, which are also very large and complex. With the development of technology and the rise of analytics, machine learning, and distributed computing, the game of cricket is evolving and will continue to evolve in the years to come, remaining a competitive advantage for teams, governing bodies, broadcasters, and fans all over the world.

7. REFERENCES

S. Naha, "'The Greatest Permanent Link between this Country and the East': Cricket Diplomacy at the End of the British Empire in South Asia, c. 1932–1950," *The English Historical Review*, vol. 140, no. 604-605, pp. 747-776, 2025.

R. V. November, J. Ras, M. S. Taliep, H. Cai, C. Nyirenda, and L. L. Leach, "The Determinants of Success in One Day International (ODI) and Twenty20 (T20) Cricket Matches: A Systematic Review and Meta-Analysis," *Applied Sciences*, vol. 15, no. 19, p. 10341, 2025.

S. Abraham, "From Analysis to Action: A Strategic Overview of Cricket Analytics," *The Game Behind the Game: Analytics in Sports*, p. 64.

R. M. Galekwa, J. M. Tshimula, E. G. Tajeuna, and K. Kyandoghere, "A systematic review of machine learning in sports betting: Techniques, challenges, and future directions," *arXiv preprint arXiv:2410.21484*, 2024.

M. J. Awan *et al.*, "Cricket match analytics using the big data approach," *Electronics*, vol. 10, no. 19, p. 2350, 2021.

P. S. Rani and T. G. Madhuri Naidu, "The 42 Vs of Big Data," *Advances in Consumer Research*, vol. 2, no. 4, 2025.

A. Eassom, S. Robertson, J. Haycraft, and M. Reid, "Validation of spatiotemporal tennis ball trajectory tracking technology," *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, p. 17543371251381488, 2025.

Rayan, R. A., Tsagkaris, C., Zafar, I., Moysidis, D. V., & Papazoglou, A. S. (2022). Big data analytics for health: a comprehensive review of techniques and applications. *Big data analytics for healthcare*, 83-92..

J. Tyagi, K. Agarwal, L. Seth, and K. Khurana, "BOUNDRY BOX: Cricket Data Analytics & Fantasy Team Model," *Available at SSRN 5192677*, 2025.

G. Revathy, S. Madeshwaran, C. Chidambaranathan, and M. Sakthivel, "AI and Data Science: Transforming IPL Strategy and Success," in *Advanced AI and Data Science Applications*: Auerbach Publications, 2025, pp. 19-31.

G. H. Al Masud, R. I. Shanto, I. Sakin, and M. R. Kabir, "Effective depression detection and interpretation: Integrating machine learning, deep learning, language models, and explainable AI," *Array*, vol. 25, p. 100375, 2025.

Scientific Reports, vol. 15, no. 1, p. 29983, 2025.

A. H. Kumarji, S. Sharma, and D. Mehrotra, "Optimizing Cricket Batting Techniques Through Pose Estimation and Machine Learning," in *2025 International Conference on Decision Aid Sciences and Applications (DASA)*, 2025: IEEE, pp. 1621-1628.

I. J. Chukwudi, M. A. Rahman, A. H. Alenezi, H. Tao, and P. Pillai, "Optimized Hybrid Framework vs. Spark and Hadoop: Performance Analysis for Big Data Applications in Vehicular Engine Systems," *IEEE Access*, 2025.

M. V. Mishra, "Data Integration and Feature Engineering for Supply Chain Management: Enhancing Decision Making through Unified Data Processing," *International Journal of Advanced Research in Science, Communication and Technology*, vol. 5, no. 2, pp. 521-530, 2025.

N. L. Mnyone and N. Amin, "Distributed Big Data Architecture for Smart Transportation Using Hadoop, Kafka, and Apache Flink," 2026.

A. Khan, "Real-Time Sports Analytics Dashboard Using Kafka and Apache Flink," *International Journal of Advanced Research in Computer Science and Engineering (IJARCSE) US ISSN: 3071-0154*, vol. 2, no. 1, pp. (33-42), 2026.

E. Priyavadhana, T. Balakrishnan, and K. Takeaways, "BIG DATA ANALYTICS," *AGRICULTURAL SCIENCES*, p. 81.

P. Thariyan and T. Thomas, "The Women's Premier League-A Paradigm Shift in Indian 'Women's Cricket'," *Journal of Advanced Research and Innovation*, vol. 1, no. 4, pp. 1-7, 2025.

A. Butterworth, "Emerging technology and interactive feedback," in *Professional practice in sport performance analysis*: Routledge, 2023, pp. 19-46.

A. Choudhary and S. Pamidimokkala, "A Comprehensive Review of Natural Language Processing Impact on Diagnosis and Decision-Making in Healthcare, Business, Education, and Sports."

M. Waldron and P. Worsfold, "Differences in the game specific skills of elite and sub-elite youth football players: Implications for talent identification," *International journal of performance analysis in sport*, vol. 10, no. 1, pp. 9-24, 2010.

M. Bhatnagar and M. Bhatnagar, "Analyzing key factors influencing IPL cricket scores using explainability and multimodal data," *Journal of Quantitative Analysis in Sports*, vol. 21, no. 3, pp. 253-267, 2025.

M. Goel, *AI based Sports education System: AI based Sports education System*. Geh press, 2026.

J. Yuan, Q. Zeng, A. Wang, Y. Zhang, and J. Li, "Machine Learning Applications in Non-Contact Lower Limb Sports Injury Prediction: A Systematic Review," *Journal of Sports Science and Medicine*, vol. 25, no. 1, pp. 172-194, 2026.

R. Trivedi, "Revolutionizing Fan Engagement: Data Analytics in Modern Sports," in *Sports Data Analytics: Techniques, Applications, and Innovations*: Springer, 2026, pp. 231-242.

Y. Agrawal and K. Kandhway, "Predicting the winner of a Twenty20 international cricket match: classification and explainable machine learning approach," *The Journal of Prediction Markets*, vol. 18, no. 1, pp. 47-64, 2024.

M. S. Ayub *et al.*, "CAMP: A context-aware cricket players performance metric," *Journal of the Operational Research Society*, vol. 75, no. 6, pp. 1140-1156, 2024.

S. Viswanadha, K. Sivalenka, M. G. Jhawar, and V. Pudi, "Dynamic Winner Prediction in Twenty20 Cricket: Based on Relative Team Strengths," in *MLSA@ PKDD/ECML*, 2017, pp. 41-50.

W. Ahmed, "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.

M. Yasir, L. Chen, S. A. Shah, K. Akbar, and M. U. Sarwar, "Ongoing match prediction in T20 International," *International Journal of Computer Science and Network Security*, vol. 17, no. 11, pp. 176-181, 2017.

F. Munir, M. K. Hasan, S. Ahmed, and S. Md Quraish, "Predicting a T20 cricket match result while the match is in progress," Brac University, 2015.

S. Chakraborty *et al.*, "Cricket data analytics: Forecasting T20 match winners through machine learning," *International Journal of Knowledge-based and Intelligent Engineering Systems*, vol. 28, no. 1, pp. 73-92, 2024.

P. Sahu and J. K. Mantri, "Performance evaluation of data mining and neural network based models for diabetes prediction," in *2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, 2023: IEEE, pp. 1-10.

T. P. Dalal, H. Shah, T. Kanjariya, and D. Joshi, "Cricket match analytics and prediction using machine learning," *International Journal of Computer Applications*, vol. 186, no. 26, pp. 27-33, 2024.

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, p. 100025, 2022.

K. Dhinakaran and S. Anbuchelian, "Enhanced cricket match prediction using kernel methods for feature extraction and back-propagation neural networks," *Scientific Reports*, 2026.

Rayan, R. A., Zafar, I., & Tsagkaris, C. (2021). Artificial intelligence and big data solutions for COVID-19. In *Intelligent data analysis for COVID-19 pandemic* (pp. 115-127). Singapore: Springer Singapore..

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, 2024.

R. P. Bunker and F. Thabtah, "A machine learning framework for sport result prediction," *Applied computing and informatics*, vol. 15, no. 1, pp. 27-33, 2019.

J. Yang, "Machine Learning-Based Injury Risk Prediction Models for Athletes in Rehabilitation," *J. Comb. Math. Comb. Comput*, vol. 127, pp. 9385-9406, 2024.

N. M. Billa, P. Akki, S. Kaliappan, M. Radha, G. G. Priya, and S. K. Pathuri, "A ML Approach for Predicting the Winner of IPL-24 Using Novel-Hybrid Classifier," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 2024: IEEE, pp. 1-6.

M. Saeed *et al.*, "Harnessing Deep Learning Models for Guide RNA Optimization and Off-Target Prediction in CRISPR Systems," *Biotechnology Journal*, vol. 21, no. 6, p. e70255, 2026, doi: https://doi.org/10.1002/biot.70255.

M. S. Khan *et al.*, "Next-Generation Artificial Intelligence Strategies for Mechanistic Cancer Target Discovery and Drug Development: A State-of-the-Art Review," (in eng), *Int J Mol Sci*, vol. 27, no. 9, Apr 30 2026, doi: 10.3390/ijms27094028.

T. Talaei Khoei, "Sports Analytics Foundations: From Data Collection to Athlete Management Systems," *Sports Data Analytics: Techniques, Applications, and Innovations*, pp. 1-30, 2026.

M. P. R. Mahin, E. Ara, and R. U. Islam, "Cricket Analytics: ODI Match Score Prediction and Resource Metric Modeling using Machine Learning Model," in *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, 2024: IEEE, pp. 1-6.

M. Verma and U. Sharma, "Prediction and comparison of compressive strength of geopolymer concrete by using decision tree and random forest regression model," *Asian Journal of Civil Engineering*, vol. 27, no. 1, pp. 339-354, 2026.

R. Ahamad, K. N. Mishra, A. A. Jiman, S. K. Ray, and P. N. Barwal, "A Data-Driven Approach for Predicting Cricket Innings Scores Using Intelligent Machine Learning Model," in *2026 International Conference on AI Innovations and Industry (ICAIII)*, 2026: IEEE, pp. 1-8.

D. Selvamuthu and D. Das, "Descriptive statistics," in *Introduction to probability, statistical methods, design of experiments and statistical quality control*: Springer, 2024, pp. 201-240.

S. D. Roy and S. J. Jackson, "Sport and Corporate Nationalism in India," *Sport, Advertising and Global Promotional Culture: Identities, Commodities, Spaces and Spectacles*, 2025.

K. Gkanatsios, "Analysis of transition offenses between winning and losing teams," Lietuvos sporto universitetas., 2025.

K. Ganatra, A. Kakkad, and C. Shingadiya, "Machine Learning for Sports Score and Winner Prediction: A Critical Review of Research Gaps and Model Enhancements," *Emerging Perspectives and Applications of Computational Intelligence and Smart Systems*, pp. 333-337, 2026.

D. Krishnan, S. Anbuchelian, and P. Vanitha, "Enhancing ODI Cricket Score Forecasting Using AI: A Data-Driven Approach with Machine Learning Models," in *2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, 2025: IEEE, pp. 1-6.