SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I5P105

Volume 11 Issue 5, 44-58, May 2024
© 2024 Seventh Sense Research Group®

Original Article

EFSM-MLB: An Ensemble Feature Selection Model for Better Outcome Prediction in Major League Baseball Using Filter and Embedded Methods

Deepak Pandey[1], Rajeev Gupta[2]

[1]*M.M. Institute of Computer Technology & Business Management, Maharishi Markandeshwar (Deemed to be University), Haryana, India.*
[2]*Department of Computer Science and Engineering, M.M. Engineering College, Maharishi Markandeshwar (Deemed to be University), Haryana, India.*

[2]*Corresponding Author : rajeev.gupta@mmumullana.org*

Received: 11 March 2024 Revised: 20 April 2024 Accepted: 08 May 2024 Published: 31 May 2024

Abstract *- Major League Baseball (MLB) stands as one of the most globally renowned and widely played tournaments at the international level in the realm of sports research. Predicting the key input variables of a match in MLB based tournament is very challenging. The selection process involves choosing which variables are more important for match prediction, as teams often use Sabermetrics in feature selection for an accurate selection of players. The current study aims to identify the major input variables that influence MLB team winnings. The authors of this research suggested an ensemble feature selection model for a better and more accurate outcome of a match. The proposed mechanism is tested on an open-access dataset of major leagues from 2005 to 2023, which is freely available on Baseball-Reference. The authors implement the proposed model on a set of sixty different offensive and defensive game features. Results obtained from deep analysis and implementation using linear regression and Correlation indicate a positive or negative association with win percentage. Here, the suggested model ranks all MLB variables from highly correlated to lesser correlated variables according to their association with win percentage. Pitching characteristics are found to be more important for forecasting match outcomes in favour of winners during this feature selection process. Furthermore, it has been discovered that run differential is a major factor in match prediction.*

Keywords *- Correlation, Feature selection, Machine Learning, Major League Baseball, Regression, Run difference.*

1. Introduction

The matter of key variable selection has undergone general exploration for various objectives, including clustering, classification, and function approximation. Feature selection and Feature extraction are the two principal methods, that are employed to identify the subclass of probable input key variables. Feature selection reduces dimensionality by choosing a subset of the original input variables, whereas feature extraction transforms the original variables to generate more relevant features.

Streamlining the multitude of characteristics becomes beneficial when dealing with data containing a substantial number of features, enhancing data analysis. As the number of features grows, the accuracy of the applied learning method tends to diminish. The primary motivation behind game feature reduction is to restrict the computational time of specific learning algorithms, ultimately reinforcing their accuracy. As for Major League Baseball (MLB), the top professional baseball league in the world, eminent researchers have prompted academic interest in forecasting game outcomes, player performance, and player value. Major League Baseball (MLB), a prominent North American professional baseball league, amassed a staggering $10.32 billion in revenue in 2022, equating to an average franchise income of $344 million [1] per team. Unraveling the pivotal variables influencing game outcomes is crucial, with input variables like hitting, pitching, and fielding. These input variables play a significant role in predicting game outcomes. MLB teams invest lots of money in statistical analysis to improve their team performance. In the past, the judgment of the match accuracy was very poor because the researcher's studies were based on subjective circumstances. However, now researchers use various machine learning techniques to analyze and select the most suitable game input features with the help of open-access databases like Baseball Reference, Fan Graph, Retro sheet, and Lehman's Database. Baseball game match outcomes are challenging to predict because many factors affect the game outcome, including feature selection, team spirit, and weather.

1.1. Feature Selection

In the realm of machine learning research, the significance of feature selection is paramount for enhancing the predictive model's performance. The goal of machine learning in selection methodologies is to identify the optimal set of features, enabling the development of highly optimized models for studied phenomena. The objective is to reduce the number of input variables, not only to diminish computational costs associated with modeling but also, in some instances, to enhance overall model performance. Game features selection as a key input serves to pinpoint all inputs influencing the phenomenon of interest and serves as a crucial data pre-processing step across various domains within machine learning [2-6].

Prominent feature selection is widely applied in diverse applications, including function approximation, classification, and clustering [7]. The challenge in extracting the most relevant variables stems from the huge set of features or key input variables the inter-correlated introducing delicacy, and the availability of key input variables that lack influence on the considered phenomenon with lacking predictive power.

For the finalization of an optimized subset of key input features, the following points should be considered:

- Relevance: The number of optimized and selected features needs validation to prevent an insufficient representation that lacks meaningful information in the realm of machine learning.
- Computational Efficiency: In machine learning, the computational burden escalates with a high number of specified input variables, especially evident when employing artificial neural networks. Additionally, the introduction of duplicate and irrelevant variables poses challenges during the training of neural networks, as extraneous features contribute to noise and impede the efficiency of network training.
- Knowledge Improvement: The optimal selection of input variables contributes to a deeper understanding of the process behavior.

The objective of feature selection, a crucial stage in machine learning and data analysis, is to identify the most pertinent features or traits from the original dataset to boost interpretability, lower computational complexity, and improve model performance. Figure 1 illustrates the various existing techniques for choosing features.

1.1.1. Filter Method

Generally, feature selection is done independently of any machine learning approach using filter techniques as a pre-processing step. Rather, the selection of features is done based on how well they perform in different statistical tests that evaluate their Correlation with the end variable. Although filtering techniques are computationally efficient, overfitting problems could arise. Correlation, mutual information, Chi-square test, information gain and gain ratio, ANOVA (analysis of variance), and variance threshold are a few popular examples of filter techniques. The filter strategy's main benefit is its minimal computational cost, which guarantees the speed of the model. However, the filter technique's primary flaw is that it does not depend on the algorithm used to build or adjust the model that is given [8-10].

1.1.2. Wrapper Method

These techniques assess subsets of features by utilising a particular machine learning algorithm's prediction capability. A "wrapper method" in the context of feature reduction refers to a technique where a subset of features is selected based on how well a machine learning model performs with that subset.

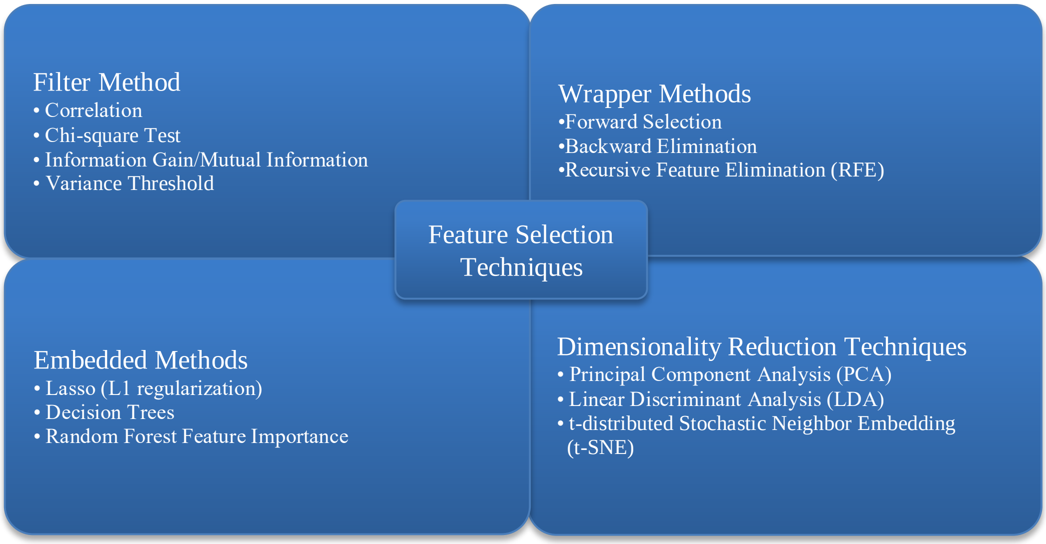
Filter Method
• Correlation
• Chi-square Test
• Information Gain/Mutual Information
• Variance Threshold
Wrapper Methods
•Forward Selection
•Backward Elimination
•Recursive Feature Elimination (RFE)
Feature Selection Techniques
Embedded Methods
• Lasso (L1 regularization)
• Decision Trees
• Random Forest Feature Importance
Dimensionality Reduction Techniques
• Principal Component Analysis (PCA)
• Linear Discriminant Analysis (LDA)
• t-distributed Stochastic Neighbor Embedding (t-SNE)

Fig. 1 Features selection techniques

Unlike filter methods, which rely on statistical measures like correlation or mutual information, wrapper methods use the performance of a specific machine learning algorithm as a criterion for feature selection. Typical wrapper methods include forward selection, backward selection, and Recursive Feature Elimination (RFE). In forward selection, the process begins with a blank set of features and adds them one at a time, and monitoring performance along the way. In backward elimination, it assesses performance at each stage as it begins with all features and eliminates them one at a time. The recursive feature elimination method uses feature weights or coefficients as a basis; the model is trained iteratively, and the least significant features are eliminated.

1.1.3. Embedded Method

Embedded methods for feature selection involve performing feature selection as part of the model training process itself. This means that feature selection is integrated into the algorithm's learning process rather than being performed as a separate step before or after training. Embedded methods are particularly common in algorithms that inherently perform feature selection or regularization during training. It is a method of feature selection that is integrated into the process of training a machine learning model, and these techniques automatically select relevant features based on the inherent properties of the model. Some common techniques of embedded method are Lasso (Least absolute shrinkage and selection operator), Tree based method (e.g. random forest, gradient boosting), Elastic net (combination of (Lasso and Ridge), Regularized linear models (e.g. Ridge regression), XGBoost etc. Embedded methods are more advantageous as compared to other methods because they streamline the feature selection process within the model training phase, avoiding the need for separate feature selection steps.

1.1.4. Dimensionality Reduction Techniques

Dimensionality reduction techniques are methods used to reduce the number of input variables or features in a dataset while preserving the essential information. These techniques are particularly useful for high-dimensional datasets, where the number of features is large relative to the number of samples. Dimensionality reduction can help simplify the data, alleviate the curse of dimensionality, improve computational efficiency, and often enhance the performance of machine learning models. These techniques convert the initial feature space into a lower-dimensional space while keeping most of the pertinent data. Common techniques are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and t-distributed Stochastic Neighbor Embedding (t-SNE).

Numerous factors, such as the dataset's nature, the intended interpretability level, computing limitations, and the machine learning algorithm being used, influence the choice of feature selection approach. To determine which attributes are most pertinent for a certain activity, it is common practice to combine various approaches and engage in iterative experimentation.

1.2. MLB Input Variable

In the context of Major League Baseball (MLB), input variables typically refer to the various features or metrics used in statistical analysis, scouting, or performance evaluation. These variables can be diverse and cover different aspects of the game. Generally, MLB input variables are classified into three categories: Batting, Pitching and Fielding variables. Input variables generally refer to the data points or features that are used as inputs to predictive models, statistical analyses, or machine learning algorithms. In the case of MLB analysis, these could include any of the metrics or features mentioned earlier, depending on the specific analysis or prediction task at hand.

Table 1. MLB batting / hitting variable

Batting/Hitting Variable (Features)	Abbreviation
Plate Appearances	PA
Runs/Scored	R
Hits	H
Doubles Hit	2B
Triples Hit	3B
Home Run Hit	HR
Runs Batted In	RBI
Stolen Bases	SB
Caught Stealing	CS
Bases on Balls (Walks)	BB
Sacrifice Hits	SH
Sacrifice Flies	SF
Batting Avg. on Balls in Plays	BABIP
Grounded Ball Percentage	GB%
Left on Base	LOB
Strikeouts	SO
Batting Average	BA
On Base Percentage	OBP
Slugging Percentage	SLG
On Base Plus Slugging	OPS
Times Hit by a Pitch	HBP
Strikeouts	SO
Batting Average	BA
On Base Percentage	OBP
Slugging Percentage	SLG
Intentional Bases on Balls	IBB
Isolated Power	ISO
Based on Balls to Strike Out Ratio	BB/K
Fly Ball Percentage	FB%

1.2.1. Batting Variable

Batting is the act of coming up against the pitcher of the other side and attempting to hit the baseball ball. A batter, also known as a hitter, is the player who swings his bat to hit the ball. The common batting variables are Runs Batted in (RBI), On-Base Percentage (OBP), Slugging (SLG), Batting Average (BA), hit and home run. The detailed list of MLB batting variables is mentioned in Table 1.

1.2.2. Pitching Variable

The act of throwing the baseball towards home plate to begin a play in baseball is known as pitching. The common pitching variables are Earned Runs (ER), Strikes (ST), Earned Run Average (ERA), and Wild Pitches (WP). The detailed list of MLB pitching variables is mentioned in Table 2.

Table 2. MLB pitching variable

Pitching Variable (Features)	Abbreviation
Earned Run Average	ERA
Saves	SV
Innings Pitched	IP
Hits/Hits Allowed	H
Runs Allowed	R
Home Runs Hit/Allowed	HR
Bases on Balls/Walks	BB
Intentional Bases on Balls	IBB
Times Hit by a Pitch	HBP
Strikeouts per 9 innings	SO/9
Runners Left on Base	LOB
Number of Pitches in the PA	Pit
Strikes	Str
Balks	BK
Wild Pitches	WP
Batters Faced	BF
Fielding Independent Pitching	FIP
Walks and Hits per Inning Pitched	WHIP
Hits Allowed per 9 Innings Pitched	H/9
Home Runs per Nine Innings	HR/9
Bases on Balls per 9 Innings Pitched	BB/9
Balks	BK
Ground Balls Percentage	GB%
Fly Balls Percentage	FB%
Line Drive Percentage	LD%
Balls	B

1.2.3. Fielding Variable

Fielding holds immense significance in baseball, as defensive players aim to prevent runs and secure outs, ultimately allowing their team to take their turn at bat. The act of fielding involves seizing the ball and strategically delivering it to another defensive player to thwart base runners. A fundamental metric in baseball statistics is fielding percentage, also known as fielding average. The calculation involves the sum of putouts and assists divided by the total chances. The detailed list of MLB fielding variables is mentioned in Table 3.

Table 3. MLB fielding variable

Fielding Variable (Features)	Abbreviation
Assist	A
Double Play	DP
Errors	E
Fielding Percentage	FP
Putout	PO
Deficiency Efficiency	DefEff

The key contributions of the paper can be summarized as follows:

1.3. Employment of Ensemble Feature Selection Model (EFSM-MLB)

The paper introduces an innovative Ensemble Feature Selection Model (EFSM-MLB) specifically tailored for Major League Baseball (MLB) games.

- EFSM-MLB is designed to enhance precision in the predictive process, thereby elevating the overall efficacy of MLB game predictions.
- By leveraging ensemble techniques, EFSM-MLB discerningly selects pivotal features while eliminating redundant and inconsequential ones, leading to heightened prediction power.

1.4. Feature Selection Approach Based on Correlation Analysis

The paper presents a novel feature selection approach that involves analyzing the Correlation of each variable with the team winning in MLB games.

- Through comprehensive correlation analysis, the paper ranks the variables based on their Correlation with team winning, from higher to lower significance.
- This approach provides valuable insights into the key factors influencing the outcomes of MLB matches and guides the selection of influential features for predictive modeling.

1.5. Enhancement of Prediction Model Accuracy, Interpretability, and Flexibility

The proposed EFSM-MLB model significantly improves the accuracy, interpretability, and flexibility of prediction models for MLB matches.

- By effectively selecting key input variables, EFSM-MLB enhances the predictive power of the model, leading to more accurate and reliable predictions.
- Additionally, the model's interpretability is enhanced through the transparent selection process of pivotal features, enabling better understanding and insights into the factors driving game outcomes.

Furthermore, the flexibility of the prediction model is increased, allowing for adaptation to different scenarios and conditions in MLB games.

The subsequent sections of the paper are structured as follows: Section 2 presents the related work on MLB match variables. Section 3 presents the materials and methods used in this study, and Section 4 discusses the result analysis comparative state-of-art. Finally, in Section 5 conclusion and future research directions are presented.

2. Related Work

This section introduces the related literature based on variables used in MLB matches with or without feature selection. Numerous research works have investigated feature selection models for Major League Baseball (MLB) result prediction. Li (2022) outperformed earlier research by utilizing a Support Vector Machine (SVM) with feature selection to reach a prediction accuracy of about 65% [10]. In order to increase prediction accuracy, Huang (2021) additionally employed feature selection; an Artificial Neural Network (ANN) and SVM produced the best results. Hoang (2015) [11] concentrated on pitch prediction and employed feature analysis and selection to achieve a moderate improvement. Sidle (2017) expanded the binary pitch prediction challenge to a multi-class problem, investigating the application of adaptive feature selection techniques together with SVM, bagged random forests, and linear discriminant analysis [12-13]. All of these experiments demonstrate how feature selection methods can improve MLB outcome prediction. Chen et al. (2014) utilized logistic regression for feature selection, specifically focusing on input variables from starting pitchers [14]. Soto Valero (2016) employed five feature selection approaches in Weka, utilizing a majority vote procedure to select 15 crucial features from an initial pool of 60 variables [15]. Trawinski (2010) explored eight feature selection methods in Weka to identify the most valuable attributes from a set of 15 variables. Previous studies have employed diverse feature selection approaches with the common objective of minimizing the number of features and selecting the most promising variables to enhance prediction accuracy [16-18]. Table 4 demonstrates the literature review based on variables used in MLB matches with feature selection.

Table 4. Literature review based on variables used in MLB matches with feature selection

Author(s)	Type of Variables	No. of Variables	Input Variable	Feature Selection Method Used	Dataset	Methodology	Observations
Shu-Fen Li Et.al 2022	Hitting, Pitching	12	RBI, SO, LOB, H, BB, H, ER, Win %, R, H, OBP, OPS,	Wrapper Method, RFE (Recursive Feature Elimination)	• Baseball-Reference • MLB game data of 30 teams for the 2015-2019 seasons • 4 Years data	The methodology involved collecting MLB game data from 2015 to 2019 for 30 teams, splitting the dataset, using various prediction methods, conducting feature selection with RFE, and evaluating performance based on AUC and accuracy.	The prediction model exclusively relies on a Single Feature Selection method, suggesting the need for diverse approaches in variable selection. Accuracy: 65% Prediction method: Support Vector Machine (SVM)
Huang and Li 2021	Hitting, Pitching, Home vs away	25	GSC, FB, HR, H/A, IR, H, AB, BA, BB, PA, ERA, SLG,	Filter method, Relief	• Baseball-Reference • MLB game data of 30 teams during the	The methodology involved collecting data from the 2019 MLB season, normalizing the	The effectiveness of prediction models under different feature selection

			OBP, STR, PIT, OPS, IP, PO, BF, PIT, STR, CTCT, IS, H, BB		2019 season. • 1 Year data	data, performing feature selection, and evaluating the prediction accuracy using fivefold cross-validation.	approaches is yet undetermined. Accuracy: 94.18% Prediction method: Artificial Neural Network (ANN)
Andrew Y. Cui 2020	Hitting, Pitching	9	ISO, FIP, HR/9, K/BB, K/9, WHIP, OBP, ELO, RDBG	Embedded method	• RetroSheet Game Logs and Lahman database • MLB game data of 30 teams during the 2000-19 season. • 19 Year data	The methodology involved collecting MLB game data from 2000 to 2019 for 30 teams, splitting the dataset, performing feature selection, and evaluating the prediction accuracy.	Original features are very less, and only one features selection method is used. Accuracy: 61.77% Prediction method: Logistic Regression (LR)
Soto Valero 2016	Batting, Fielding, Pitching, Sabermetrics Statistics	15	PE, WP, RC, Home Won Prev, Visitor Won Prev, BABIP, FP, Pitcher A, OBP, SLG, Visitor League, Home Versus Visitor, Stolen, Is Home Club, Log5,	Filter Method, SignificanceAttributeEval, ChiSquaredAttributeEval, Correlation AttributeEval, GainRatio AttributeEval, ReliefF AttributeEval	• RetroSheet Game Logs and Lahman database • MLB game data of 30 teams during the 2005 - 2014 season. • 9 Year data	The methodology involved using sabermetrics statistics and four data mining methods to predict MLB game outcomes, utilizing nine years of past data and employing stratified 10-fold cross-validation to assess predictive capabilities.	Valero employed filter technique methods. The performance of prediction models using regression techniques and wrapper approaches remains unknown. Accuracy: 58.92% Prediction method: Support Vector Machine (SVM)
Chen et al. 2014	Batting, fielding, Pitching	8	Game score(H), SO(A), Earned run(A), Strike out(H), Base on balls(A),	Embedded Method	• Baseball-Reference • MLB game data of 02 teams during the 2006 - 12 season.	The methodology comprised gathering MLB game data for two teams from 2006 to 2012, dividing the dataset,	The logistic technique is used for feature selection, and SP (starting pitcher) was chosen as an input variable.

			BB(A), SO(A), WHIP(H)		• 6 Year data	choosing features, and assessing the prediction accuracy.	Accuracy: 72.22% Prediction method: Artificial Neural Network (ANN)
Jia et.al 2013	Batting, Pitching	7	RBI, H, E, BA, ERA, OBP and Win% for each team	Wrapper Method	• Baseball-Reference • MLB game data of 30 teams during the 2007 - 2012 season. • 5 Year data	The methodology comprised gathering MLB game data for 30 teams from 2007 to 2012, dividing the dataset, choosing features, and assessing the prediction accuracy.	Expanding the scope of feature selection has the potential to enhance prediction accuracy. Accuracy: 59.60% Prediction method: Support Vector Machine (SVM)

2.1. Research Problem

Some possible research gaps identified in the provided text are:

- There is a compelling requirement for additional research focusing on the integration of feature selection methods with artificial intelligence technologies for the selection of feature subsets in Major League Baseball (MLB). This paper addresses this research gap by introducing an Automated Feature Selection Mechanism, EFSM-MLB, contributing to the advancement of knowledge in this domain.
- There is a call for further investigation into the pragmatic implementation of feature selection methods, with a specific emphasis on regression and correlation techniques. Although the study's experimental method yields promising outcomes, there is a recognized opportunity to enhance the automatic selection of variables correlated with win percentage. To augment feature selection, the adoption of the EFSM-MLB model is considered for improvement.

3. Materials and Methods

This portion goes over the materials and methodologies utilized to create the model. Section 3.1 explains the basic architecture and the algorithm of the proposed ensemble feature selection model. Data collection and pre-processing are presented in Section 3.2.

3.1. Basic Architecture of EFSM-MLB Model

EFSM-MLB stands for Ensemble Feature Selection Model for Major League Baseball. This model aims to improve outcome prediction in Major League Baseball (MLB) by employing a combination of filter and embedded feature selection methods.

In machine learning, feature selection is the process of choosing a subset of relevant features or variables to use in model construction. This is done to improve model performance, reduce overfitting, and enhance interpretability. Filter methods evaluate the relevance of features independently of the model, while embedded methods incorporate feature selection directly into the model training process. The "Ensemble" aspect of EFSM-MLB suggests that it may utilize multiple feature selection techniques or models to arrive at a more robust set of selected features. Ensemble methods often combine the predictions of several base estimators to improve generalizability and robustness over a single estimator.

By applying both filter and embedded feature selection techniques specifically tailored for MLB data, EFSM-MLB aims to enhance the accuracy of outcome predictions in Major League Baseball games. This could potentially lead to better player performance analysis, team strategy development, and informed decision-making by coaches, analysts, and team management within the realm of professional baseball. Figure 2 illustrates the basic architecture of the proposed EFSM-MLB model. The main objective of this study is to select a feature subset of MLB matches from the season (2005-2023). To achieve this, the authors are using correlation and regression techniques. The algorithm of the proposed EFSM-MLB model is displayed as follows.

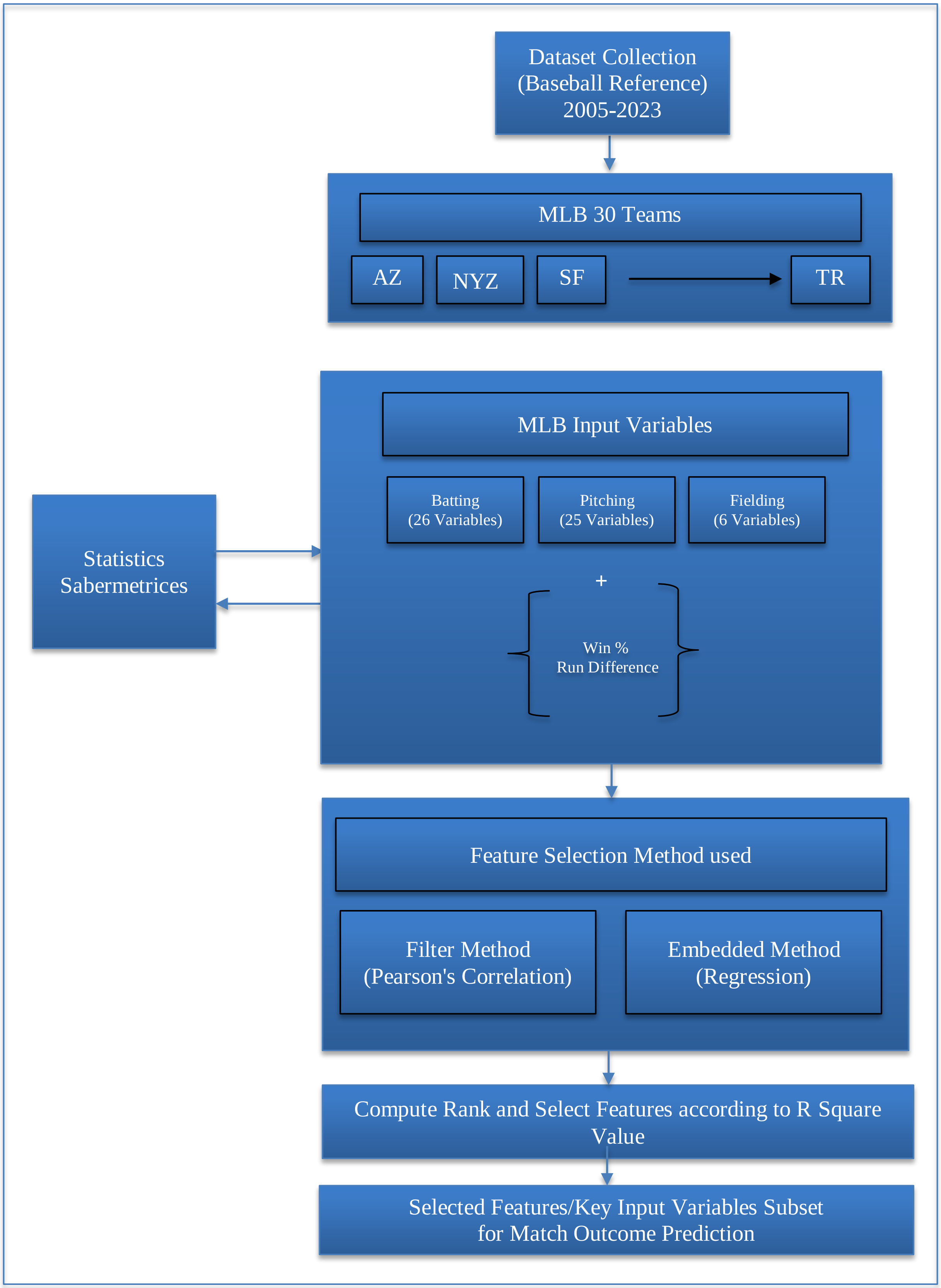

Fig. 2 Basic architecture of the proposed EFSM-MLB model

```
Algorithm: Ensemble Feature Selection Model (EFSM-MLB) for Major League Baseball matches
Input: Major League Baseball match Team data (batting, pitching and fielding)
Output: Key Input Variables with their Rank value (sorted, highest to lowest)
Begin
Step 1: Choose a data file (in .xlsx/ .CSV form) from the dataset
Apply steps to 6 and compute the average Correlation of each variable (batting, pitching and fielding) with win % and filter the features set
Step 2: Arrange Raw Data
Step 3: Calculate Correlation Matrix
Step 4: Compute the High Correlation Pairs
Step 5: Remove redundant features
Step 6: Sort the values fetched (highest to lowest order) and filtered features set on the basis of the variable’s correlation value against win%

Apply steps to 12 and evaluate the R-squared value of each variable (batting, pitching and fielding) with win % and filter the features set
Step 7: Arrange Raw Data
Step 8: Initialize a Data Frame to store the results
        results_df = pd.DataFrame(columns=['Variable',
                     'R-squared Value'])
Repeat Step-9 and Step-10 through each input variable for feature in feature_columns
Begin
   Step 9:  Fit the model
         X = sm.add_constant(df[feature])
         y = df['win %']
        model = sm.OLS(y, X).fit()
   Step 10: Append results to the DataFrame
End
Step 11: Display the table of R-squared values
Step 12: Sort the values fetched (highest to lowest order) and filtered features set on the basis of R-squared value against win%
Step 13: Display Key Input Variables with their Rank value (sorted, highest to lowest)
End
```

3.2. Data Collection and Preprocessing

Data collection and pre-processing in the context of Major League Baseball (MLB) involves gathering and preparing data related to baseball games, players, teams, and other relevant factors. For feature selection, thirty teams' MLB game data from the 2005–2023 season was gathered and examined. Over the 19 years specified, each squad plays about 162 games. The comprehensive dataset was sourced exclusively from Baseball-Reference, a reputable platform known for recording detailed information on batting, pitching, fielding, and game outcomes. In adherence to the ethical guidelines outlined in the Belmont Report, it is essential to highlight that all the data utilized in this study is publicly available. Consequently, there is no obligation to seek informed consent from the participants, given the public nature of the data.

Various websites recorded MLB game data, with nuances in the recorded variables. Notably, platforms like Retrosheet and Lehman database capture original game data, which can be subsequently transformed into Sabermetrics. The Baseball-Reference website, specifically accessed on 12 April 2024, offers user-friendly access to Sabermetrics and stands out for its ease in searching for a particular MLB game player or match-related data. Therefore, the authors selected the Baseball-reference website to collect team standard data related to pitching, fielding and hitting in this study. The data obtained from Baseball-Reference is categorized into team hitting, team fielding, and team pitching.

The primary focus of this research is to analyze the Correlation between various variables and a team's success in terms of winning. Prioritizing key variables like Runs scored (Run), Runs Batted In (RBI), Run Difference (RD), and Winning Percentage (Win%) is a solid approach, as these factors are closely tied to the outcome of games. Calculating the winning percentage (Win%) using Equation 1 and determining the Run Difference (RD) using Equation 2 provides a structured method for quantifying these important metrics.

Equation-1:

$$Win\% = \frac{Total\ number\ of\ matches\ win\ by\ a\ Team}{Total\ number\ of\ matches\ played\ by\ a\ Team\left(N\ matches\right)}$$

Equation-2:

$$Run\ difference\ (RD) = (Runs\ Scored - Runs\ Earned)$$

This systematic ranking based on Correlation can help identify which variables have the strongest impact on a team's success, providing valuable insights for coaches, analysts, and decision-makers in the realm of sports management and strategy.

The proposed model is a systematic method for analyzing variables that contribute to team success in sports, particularly baseball. Using run difference as a metric is a common practice in baseball analysis, as it provides a simple yet insightful measure of a team's performance. The run difference, calculated by subtracting runs allowed from runs scored, serves as a straightforward metric to gauge a team's performance. It condenses the team's offensive and defensive capabilities into a single numerical value.

The relationship between run difference and wins is intuitive – generally, teams with higher run differences tend to win more games. This metric serves as a concise summary of a team's scoring proficiency relative to its opponents, offering a clear perspective on its performance level. It is a valuable tool for evaluating a team's competitiveness and success within the context of their league or competition.

4. Results and Discussion

The proposed model is implemented and tested in the environment mentioned in Table 5.

Table 5. Experimental environment setup

Experiment Parameter	Parameter Value
Tool	Jupyter Notebook (Python 3 – iypkernel)
Data Set	Baseball-Reference – a repository of baseball statistics for every player of MLB
Duration	19 years (2005 – 2023)
Teams	30 Teams
No. of Games	162 per season for each of the 30 teams Total: 2430 per season
No. of Input Variables	60

In this study, sixty diverse offensives, defensive, and pitching statistics from Baseball Reference are analyzed for all thirty MLB teams, comparing each metric to the respective team's win percentage across the 2005-2023 MLB seasons. The resulting average correlations (Table-6) as well as average R^2 (R-Square) (Table-7) are categorized as follows: less than 0.290 (Red) denoting no correlation, 0.291-0.500 (Blue) indicating moderate Correlation, and 0.501-1.000 (Green) signifying a strong correlation.

Table 6. Average correlation with reference to win%

Rank	Variable	Correlation	Rank	Variable	Correlation
1	win %	1.000	31	Str(P)	-0.028
2	RD(D)	0.932	32	A(F)	-0.088
3	SV(P)	0.651	33	SH(H)	-0.104
4	R(H)	0.629	34	FB%(P)	-0.108
5	RBI(H)	0.627	35	3B(H)	-0.127
6	OBP(H)	0.591	36	CS(H)	-0.138
7	OPS(H)	0.587	37	BK(P)	-0.156
8	BB/K (H)	0.532	38	HBP(P)	-0.177
9	SLG(H)	0.530	39	SO(H)	-0.183
10	PO(F)	0.529	40	DP(F)	-0.184
11	IP(P)	0.528	41	WP(P)	-0.207
12	DefEff(F)	0.503	42	GB%(H)	-0.234
13	BB(H)	0.483	43	IBB(P)	-0.236
14	ISO(H)	0.475	44	LD% (P)	-0.298
15	PA(H)	0.475	45	EC(F)	-0.350
16	HR(H)	0.453	46	Pit(P)	-0.363
17	SO9(P)	0.445	47	LOB(P)	-0.433
18	BA(H)	0.384	48	HR(P)	-0.481
19	H(H)	0.360	49	Balls(P)	-0.489
20	Fld%(F)	0.355	50	Babip(P)	-0.497
21	FB%(H)	0.306	51	HR9(P)	-0.504
22	LOB(H)	0.298	52	BB(P)	-0.532
23	IBB(H)	0.291	53	BB9(P)	-0.560
24	SF(H)	0.262	54	BF(P)	-0.593
25	GB%(P)	0.223	55	H/9(P)	-0.623
26	2B(H)	0.216	56	H(P)	-0.630
27	HBP(H)	0.186	57	FIP(P)	-0.696
28	Babip(H)	0.089	58	WHIP(P)	-0.760
29	SB(H)	0.044	59	ERA(P)	-0.767
30	LD%(H)	0.000	60	R(P)	-0.770

Table 7 . Average R-square with reference to win%

Rank	Variable	R SQUARE	Rank	Variable	R SQUARE
1	win %	1.000	31	LOB(P)	0.210
2	RD(D)	0.870	32	BA(H)	0.174
3	R(P)	0.596	33	Pit(P)	0.163
4	ERA(P)	0.591	34	H(H)	0.152
5	WHIP(P)	0.578	35	EC(F)	0.150
6	FIP(P)	0.482	36	Fld%(F)	0.147
7	SV(P)	0.441	37	LD% (P)	0.120
8	H/9(P)	0.429	38	IBB(H)	0.113
9	R(H)	0.422	39	SF(H)	0.110
10	RBI(H)	0.417	40	FB%(H)	0.108
11	H(P)	0.408	41	LOB(H)	0.102
12	OPS(H)	0.374	42	2B(H)	0.088
13	OBP(H)	0.372	43	WP(P)	0.080
14	BF(P)	0.364	44	HBP(H)	0.078
15	BB9(P)	0.329	45	HBP(P)	0.076
16	SLG(H)	0.315	46	GB%(H)	0.074
17	BB(P)	0.299	47	IBB(P)	0.067

18	BB/K (H)	0.296	48	BK(P)	0.062
19	PO(F)	0.292	49	GB%(P)	0.061
20	IP(P)	0.291	50	DP(F)	0.060
21	HR9(P)	0.276	51	SO(H)	0.058
22	Balls(P)	0.260	52	CS(H)	0.054
23	ISO(H)	0.259	53	SB(H)	0.040
24	Babip(P)	0.258	54	A(F)	0.038
25	DefEff(F)	0.257	55	3B(H)	0.037
26	BB(H)	0.255	56	Babip(H)	0.032
27	HR(P)	0.246	57	SH(H)	0.029
28	PA(H)	0.237	58	FB%(P)	0.029
29	HR(H)	0.236	59	LD%(H)	0.023
30	SO9(P)	0.235	60	Str(P)	0.023

The graphical representation of the positive correlation with win% and negative correlation with win% is shown in Figures 3 and 4, respectively. Moreover, the graphical representation of strong and moderate R^2 values with win% and weak R^2 values with win% is shown in Figures 5 and 6, respectively.

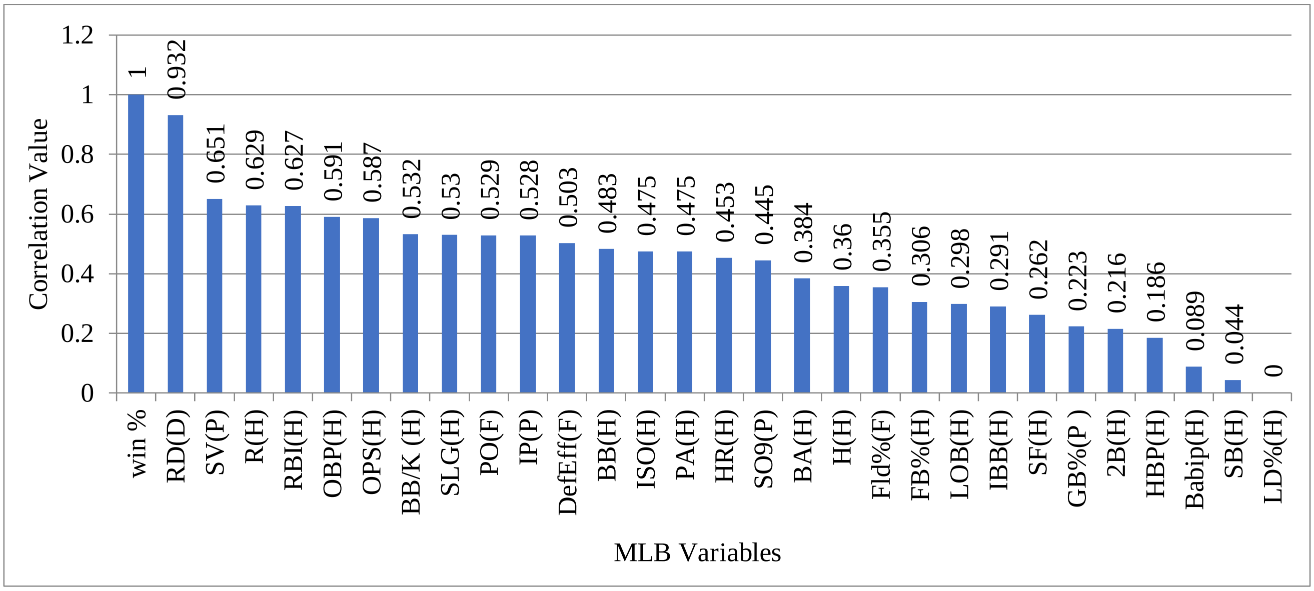

Fig. 3 Positive correlation with win%

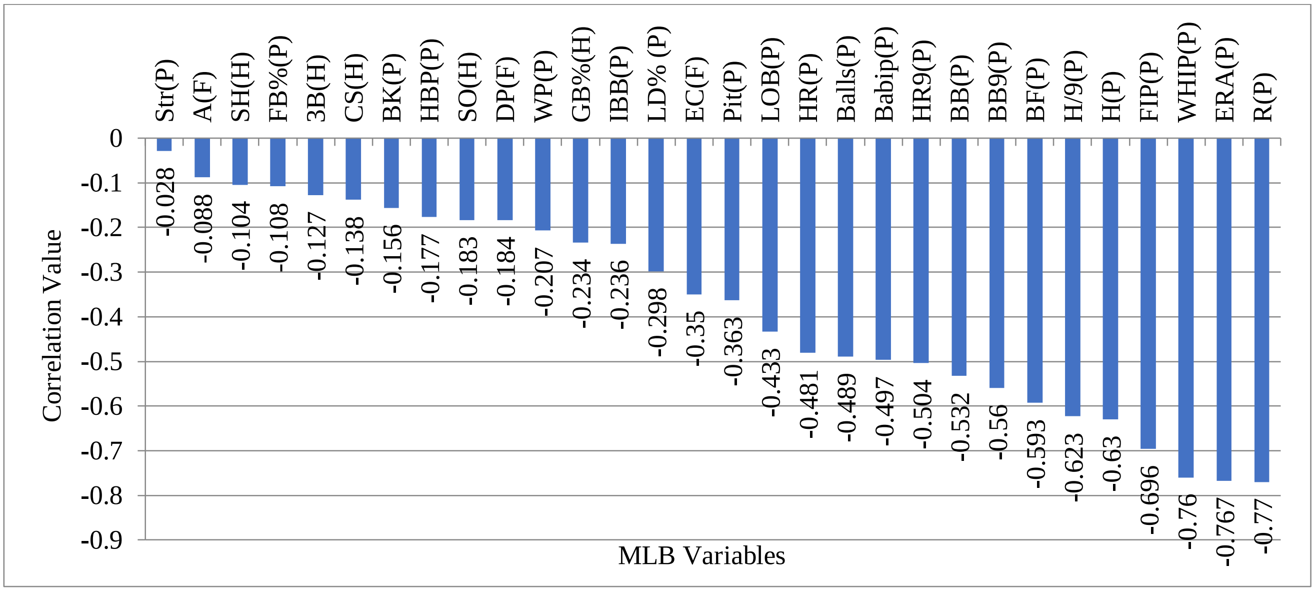

Fig. 4 Negative correlation with win%

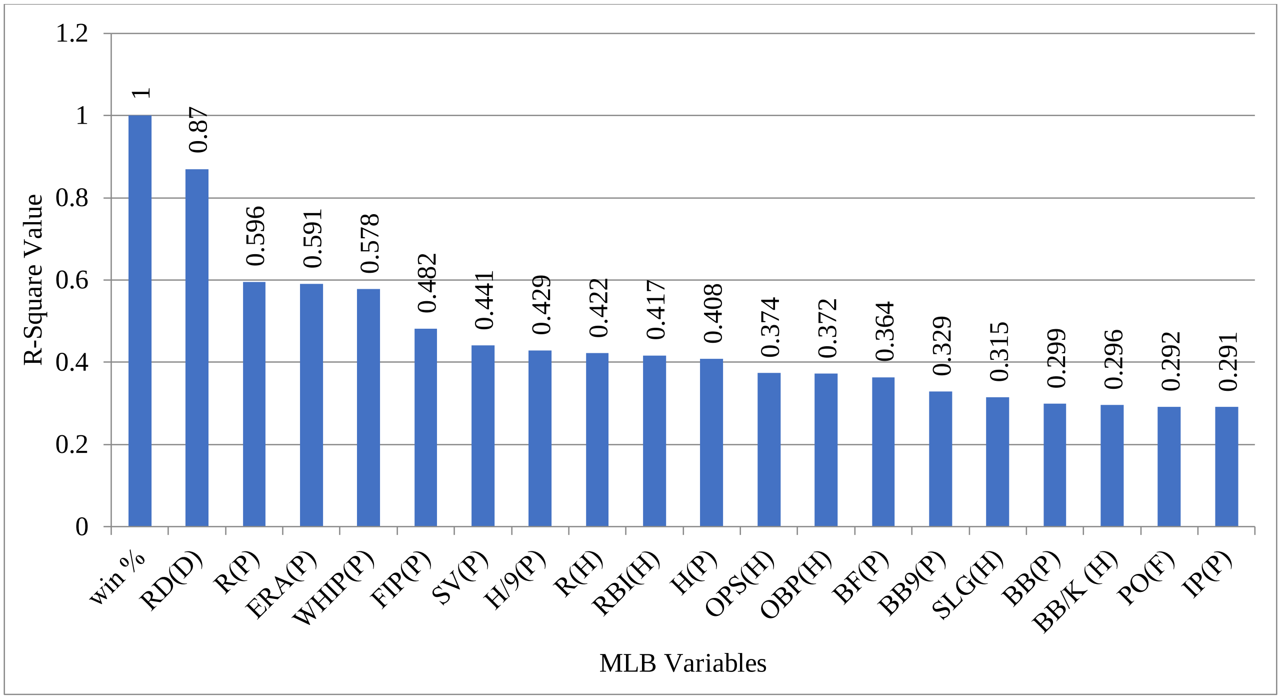

Fig. 5 Strongly and moderate R-square value with win%

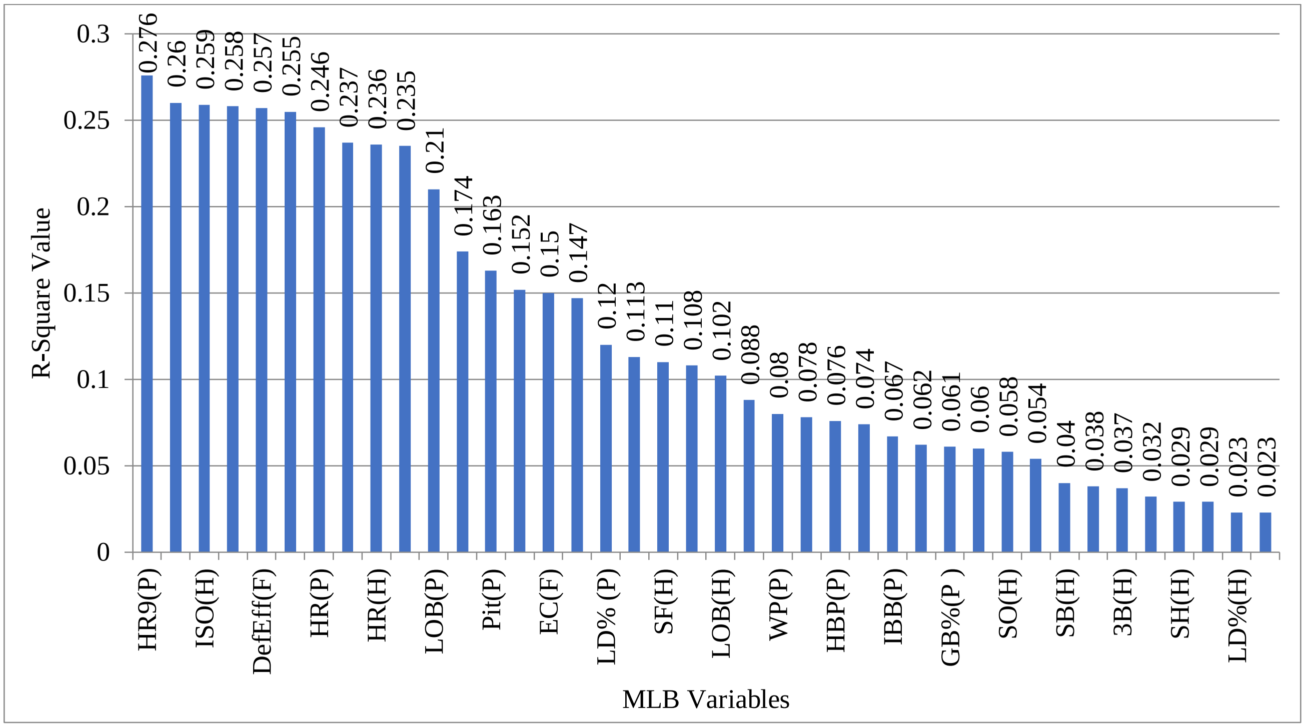

Fig. 6 Weak R-square value with WIN%

5. Comparative Analysis of EFSM-MLB with Traditional Methods

After the systematic analysis and experimental study, the proposed model EFSM-MLB identified the most relevant key variables for further prediction of the accuracy of the outcome of an MLB match. The following Table 8 shows the comparative analysis of the proposed model EFSM-MLB and traditional models in terms of the number of features used and the time duration of the data set used. Figure 7 illustrates the graphical representation of the comparative analysis of the proposed model EFSM-MLB and traditional models.

Table 8. Comparative analysis of EFSM-MLB with traditional methods

Author(s)	Type of Variables	No. of Variables	Input Variable	Feature Selection Method Used	Duration of Data Used (in Years)
Shu-Fen Li et al. 2022	Hitting, Pitching	12	RBI, SO, LOB, H, BB, H, ER, Win %, R, H, OBP, OPS,	Wrapper Method, RFE (Recursive Feature Elimination)	4
Huang and Li 2021	Hitting, Pitching, Home vs away	25	GSC, FB, HR, H/A, IR, H, AB, BA, BB, PA, ERA, SLG, OBP, STR, PIT, OPS, IP, PO, BF, PIT, STR, CTCT, IS, H, BB	Filter method, ReliefF	1
Andrew Y. Cui 2020	Hitting, Pitching	9	ISO, FIP, HR/9, K/BB, K/9, WHIP, OBP, ELO, RDBG	Embedded method	19
Soto Valero 2016	Batting, Fielding, Pitching, Sabermetrics Statistics	15	PE, WP, RC, Home Won Prev, Visitor Won Prev, BABIP, FP, Pitcher A, OBP, SLG, Visitor League, Home Versus Visitor, Stolen, Is Home Club, Log5,	Filter Method, SignificanceAttributeEval, ChiSquaredAttributeEval, Correlation AttributeEval, GainRatio AttributeEval, ReliefF AttributeEval	9
Chen et al. 2014	Batting, fielding, Pitching	8	Game score(H), SO(A), Earned run(A), Strike out(H), Base on balls(A), BB(A), SO(A), WHIP(H)	Embedded Method	6
Jia et.al 2013	Batting, Pitching	7	RBI, H, E, BA, ERA, OBP and Win% for each team	Wrapper Method	5

EFSM-MLB (Proposed Method)	Batting, Fielding, Pitching, Sabermetrics Statistics	25 (Strongly & Moderated)	RD(D), SV(P), R(H), RBI(H), OBP(H), OPS(H), BB/K (H), SLG(H), PO(F), IP(P), DefEff(F)
			BB(H), ISO(H), PA(H), HR(H), SO9(P), BA(H), H(H), Fld%(F), FB%(H), LOB(H), IBB(H)
				Filter Method
				Embedded Method
					19

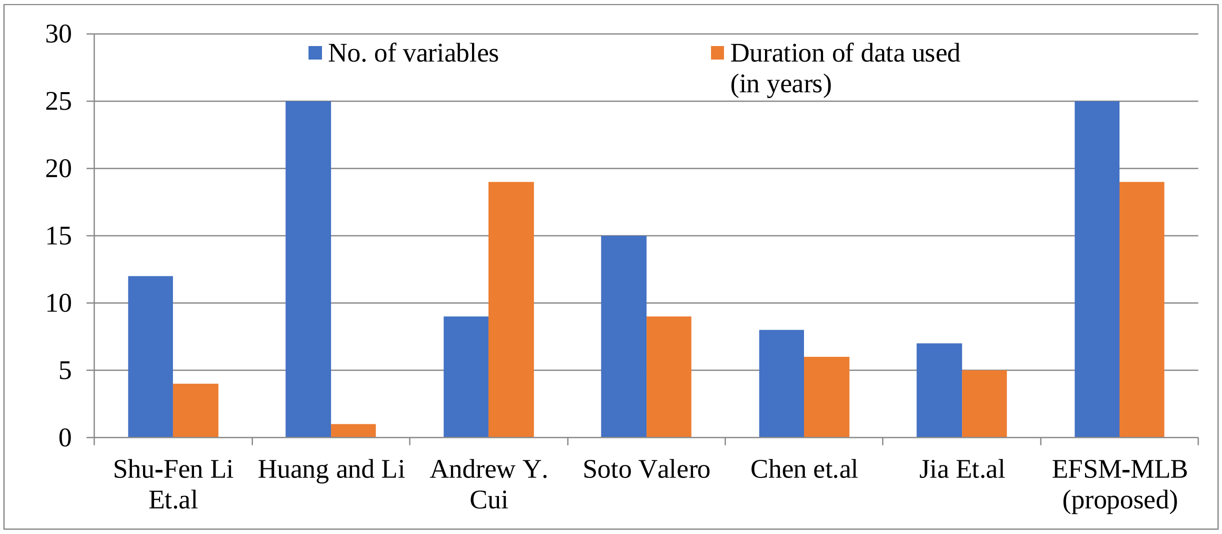

Fig. 7 Comparative analysis of EFSM-MLB with traditional methods

6. Conclusion

In summary, after a deep analysis of the Correlation between various baseball statistics and win percentages over 19 years, hitting variables emerge as a significant factor, with 06 out of 12 of the most strongly correlated variables being related to hitting. However, it is noted that there are negative correlations present, which could pose challenges. Pitching statistics also emerge as crucial, with 03 out of 05 pitching variables showing strong correlations with win percentage when considering the average R-square over 18 years. This suggests that pitching plays a pivotal role in determining a team's success. Interestingly, despite differences in correlation strength, both hitting and pitching contribute significantly to wins, with the run difference being a common factor, indicating that teams with a favourable run difference tend to win more frequently. Looking forward, the author plans to develop an AI-based model using the proposed EFSM-MLB model to enhance accuracy in predicting win percentage based on these statistics. This indicates a shift towards more advanced analytical methods to gain insights into the game of baseball.

Contribution of Author(s)

The following contributions to the work are confirmed by the authors: Design and problem conception and design: Deepak Pandey, Rajeev Gupta; data collection: Deepak Pandey; analysis and interpretation of results: Deepak Pandey; draft manuscript preparation: Deepak Pandey, Rajeev Gupta. All authors reviewed the results and approved the final version of the manuscript.

References

[1] Christina Gough, Major League Baseball Total League Revenue from 2001 to 2022, Statista, 2023. [Online]. Available: https://www.statista.com/statistics/193466/total-league-revenue-of-the-mlb-since-2005/

[2] N. Kwak, and Chong-Ho Choi, "Input Feature Selection for Classification Problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143-159, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[3] Jung-Yi Lin et al., "Classifier Design with Feature Selection and Feature Extraction Using Layered Genetic Programming," *Expert Systems with Applications*, vol. 34, no. 2, pp. 1384-1393, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[4] Silvia Cateni, Valentina Colla, and Marco Vannucci, "Variable Selection through Genetic Algorithms for Classification Purposes," *Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Applications*, pp. 1-6, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[5] Kenji Kira, and Larry A. Rendell, "A Practical Approach to Feature Selection," *Machine Learning Proceedings 1992*, pp. 249-256, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[6] Robert May, Graeme Dandy, and Holger Maier, *Review of Input Variable Selection Methods for Artificial Neural Networks*, *Artificial Neural Network - Methodological Advances and Biomedical Applications*, IntechOpen, pp. 1-28, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[7] Huanzhang Fu et al., "Image Categorization Using ESFS: A New Embedded Feature Selection Method Based on SFS," *Advanced Concepts for Intelligent Vision Systems*, Bordeaux, France, pp. 288-299, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[8] Philip Beneventano, Paul D. Berger, and Bruce D. Weinberg, "Predicting Run Production and Run Prevention in Baseball: The Impact of Sabermetrics," *International Journal of Business, Humanities and Technology*, vol. 2, no. 4, pp. 67-75, 2012. [Google Scholar] [Publisher Link]

[9] Girish Chandrashekar, and Ferat Sahin, "A Survey on Feature Selection Methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16-28, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[10] Shu-Fen Li, Mei-Ling Huang, and Yun-Zhi Li, "Exploring and Selecting Features to Predict the Next Outcomes of MLB Games," *Entropy*, vol. 24, no. 2, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Mei-Ling Huang, and Yun-Zhi Li, "Use of Machine Learning and Deep Learning to Predict the Outcomes of Major League Baseball Matches," *Applied Sciences*, vol. 11, no. 10, pp. 1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Tim Elfrink, and Sandjai Bhulai, "Predicting the Outcomes of MLB Games with a Machine Learning Approach," *Computer Science*, 2018. [Google Scholar]

[13] Randy Jia, Chris Wong, and David Zeng, "*Predicting the Major League Baseball Season*," CS 229 Machine Learning Final Projects, Autumn, pp. 1-5, 2013. [Google Scholar] [Publisher Link]

[14] Chi-Wen Chen, "Construction of the Winner Predictive Model in Major League Baseball Games: Use of the Artificial Neural Networks," *College Sports Journal*, vol. 16, no. 2, pp. 167-181, 2014. [Google Scholar]

[15] C. Soto Valero, "Predicting Win-Loss Outcomes in MLB Regular Season Games – A Comparative Study Using Data Mining Methods," *International Journal of Computer Science in Sport*, vol. 15, no. 2, pp. 91-112, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[16] Krzysztof Trawiński, "A Fuzzy Classification System for Prediction of the Results of the Basketball Games," *International Conference on Fuzzy Systems*, Barcelona, Spain, pp. 1-7, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[17] Blakeley B. McShane et al., "A Hierarchical Bayesian Variable Selection Approach to Major League Baseball Hitting Metrics," *Journal of Quantitative Analysis in Sports*, vol. 7, no. 4, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[18] Silvia Cateni, Valentina Colla, and Marco Vannucci, "Improving the Stability of the Variable Selection with Small Datasets in Classification and Regression Tasks," *Neural Process Letter,* vol. 55, no. 5, pp. 5331-5356, 2023. [CrossRef] [Google Scholar] [Publisher Link]