Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 20 April 2026
Revised 18 May 2026
Accepted 16 June 2026

Natural Resources for Human Health 2026, Vol. 6 (3s): 1118–1133

KEYWORDS:
Diabetes Mellitus, Feature Selection, Nonlinear Dynamic Conditional Relevance, Hellinger Distance, Feature Hamming Diversity, Machine Learning, Ensemble Learning, ADASYN, Explainable Artificial Intelligence.

MEFS: Multistage Ensemble Feature Selection Framework for Diabetes Risk Prediction Using Relevance, Separability, and Feature Diversity

Ashvini Sandip Patil[1], Dr. Rahul Raghvendra Joshi[2*], Dr. Milind Gayakwad[3]

[1]Research Scholar, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra, 412115, India, ashvini.patil.phd2023@sitpune.edu.in 0000-0001-8781-5357
[1]Assistant Professor, Department of Computer Science & Engineering (AI & ML), Kasegon Education Society's Rajarambapu Institute of Technology, Affiliated to Shivaji University, Sakharale, MS-415414, India, patil.ashvini04@gmail.com
[2]Associate Professor, Symbiosis Institute Technology, Pune, Symbiosis International (Deemed University), Pune, Maharashtra, 412115, India, rahulj@sitpune.edu.in, 0000-0002-5871-890X
[3]Associate Professor, Department of Information Technology, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune-411043, Maharashtra, India, mdgayakwad@bvucoep.edu.in, 0000-0002-0430-2626

ABSTRACT: Diabetes mellitus is among the most common metabolic diseases and requires early and accurate diagnosis in order to avoid any additional deterioration and complications. The performance of machine learning algorithms is promising in diabetes prediction, but their efficiency is usually hindered by redundant features, class imbalance, and poor interpretability of the developed models. Common feature selection methods usually employ only a relevance measure and neglect to incorporate nonlinear dependency, separability, and diversity, leading to poor predictive efficiency and high computational cost. To overcome these problems, we propose Multistage Ensemble Feature Selection (MEFS) framework which combines Nonlinear Dynamic Conditional Relevance Feature Selection (NDCRFS), Hellinger Distance (HD), and Feature Hamming Diversity (FHD) to select an optimal set of informative features for diabetes predictions. Firstly, the data is preprocessed using Feature-Specific Mean Variance Normalization (FSMVN) and Adaptive Synthetic Sampling (ADASYN). Then, the proposed feature selection method utilizes nonlinear analysis, distribution comparison, and redundancy elimination using an ensemble selection process. The proposed feature selection framework subsequently combines nonlinear relevance analysis, distribution-based discrimination, and redundancy reduction through an ensemble ranking strategy. Performance evaluation of the feature subset is conducted using SVM, LR, RF, XGBoost, and MLP classifiers with ten-fold cross-validation on the Pima Indians Diabetes dataset. From the experimental analysis, it can be seen that the MSFS helps in reducing the size of the original feature space by 37.5%, utilizes lesser memory with an approximation of 37% memory savings, and helps in saving computational training time without affecting prediction accuracy to any significant extent. The statistical verification process conducted through the Friedman test and Wilcoxon signed rank test proves that the suggested methodology performs as well as existing methods of feature selection without deteriorating performance.

Natural Resources for Human Health
eISSN: 2583-1194

1. INTRODUCTION

Diabetes mellitus and many other chronic diseases and infections continue to rank among the primary causes of morbidity and mortality throughout the world, driving scientific efforts aimed at improving computer algorithms that can facilitate early and precise diagnoses. The number of patients diagnosed with diabetes mellitus surpassed 537 million adults worldwide in 2021 with expected numbers going up to 783 million by 2045 [1]. Similar burdens exist for cardiovascular disease, chronic kidney disease (CKD), and many other cancers. Machine learning (ML) has emerged as a key technology for overcoming this challenge, because an ML model can reliably find intricate, non-linear connections between clinical and laboratory variables and a certain disease.

One of the challenges that remains constant to the development of reliable ML diagnostic models, is that of high dimensionality and redundancy of medical data. Data from clinical and laboratory examinations are usually gathered for patient diagnosis on a general basis and not with a certain predictive analysis objective in mind; hence, the variable set generated by such data contains many irrelevant, redundant, noisy, and weakly related features. This leads to increased computation cost and overfitting, which in turn to decrease in prediction accuracy. On the other hand, feature selection (FS) deals directly with this challenge by selecting the subset of original features that have a minimum number of features that would guarantee or enhance prediction performance.

An objective of feature selection is to find optimal feature subset which provide maximal performance in prediction and minimal redundancy and computation cost. The benefits of feature selection include higher classifier performance, lower probability of over-fitting, better understanding of the model, and faster training time. These aspects become especially valuable when it comes to healthcare applications, because doctors need to have reliable and efficient decision support system that will help in detecting relevant biomarkers. Many studies show that proper feature selection can greatly improve the performance of predicting diseases in different areas, such as diabetes, heart disease, breast cancer, Parkinson's disease and chronic kidney disease.

Feature selection methods are traditionally classified into four types. The filter type ranks and scores each individual feature on the basis of its statistical or information-theoretical association with the class label, without any use of machine learning techniques, where the associations could be estimated through different metrics like the chi-square test, mutual information, Pearson correlation, or Fisher's score. The wrapper technique uses the cross-validated performance of the selected subset of features through repeated trainings and validations of a particular classifier, which consists of classical searches such as Recursive Feature Elimination (RFE) and Sequential Forward/Backward Selection, as well as metaheuristic searches inspired by nature like Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Grey Wolf Optimization (GWO).

Despite the considerable advances made by feature selection algorithms that currently exist, there are several major areas for future research. First of all, the traditional filter approach considers each feature separately and ignores the potential presence of higher-order non-linear feature dependencies. Although wrapper approaches offer better subset optimization, they come with huge computational costs that hinder the application of such approaches to large-scale medical data sets. The embedded feature selection approach is integrated into the learning process; however, its success largely depends on the nature of the chosen learning algorithm. Despite the success of hybrid feature selection approaches which have already shown the improvements in prediction accuracy due to the combination of different feature selection approaches, most current frameworks combine two techniques and mainly concentrate on improving the classification performance. Therefore, the synergistic feature selection framework that would consider both feature relevance and feature class discrimination is mostly overlooked.

The medical data sets are characterized by different feature distributions, nonlinear associations between clinical features, and varying discrimination characteristics across disease classes. Hence, the problem of selecting informative features is not limited to relevance estimation only. An efficient feature selection methodology should be capable of performing not only the task of selecting the most relevant features but also of assessing the discriminative capability of the selected features concerning the various disease classes and eliminating structural redundancies in the selected subset of features.

Existing feature selection methods usually evaluate features using only one criterion, such as relevance, redundancy, or statistical separation. As a result, they have limited capability to identify an optimal feature subset that are relevant, diverse, and non-redundant. The relevance based feature selection techniques tend to pick highly correlated features while those based on

diversity might ignore the interdependence between the features and the target class. On the other hand, the distribution-based feature selection technique will improve class discrimination ability regardless of redundancy problem.

Imbalanced class distribution is another major issue in medical datasets. The minority class get dominated by the majority class and hence the classifier gets biased towards the majority class. There are many different approaches to overcome this problem like SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling) and hybrid approaches are used to boost the minority data samples to balance the dataset.

To deal with the above challenges, this paper proposes a Multistage Ensemble Feature Selection (MEFS) framework. The contributions made in this work are as follows:

1. Multistage ensemble-based feature selection framework that combines NDCRFS, Hellinger Distance, and Feature Hamming Diversity to achieve feature relevance, separability, and diversity at the same time.
2. Preprocessing methodology that includes feature specific mean variance normalization (FSMVN) and ADASYN for achieving consistency of features and dealing with the issue of class imbalance before feature selection.
3. Extensive experiments with five state-of-the-art classifiers, such as Logistic Regression, SVM, Random Forest, XGBoost, and Multi-Layer Perceptron, under ten-fold cross-validation setting.
4. Statistical significance assessment using Friedman and Wilcoxon signed-rank tests.

The remainder of this paper is organized in the following manner. In Section 2, the review of literature concerning diabetes prediction and present day methods of feature selection is carried out. In Section 3, our proposed multistage feature selection method is discussed. The experimental setup and methodology of this work are presented in Section 4. The discussion of experimental

is provided in Section 5. Conclusions and future work are summarized in Section 6.

2. LITERATURE REVIEW

Filter-based approaches. An extensive set of studies uses the univariate statistical filtering approach for selection of the predictive features for diabetes disease. In particular, the use of chi-square filtering approach on the PIMA data in combination with SMOTE balancing was studied by Giri et al. [6] using six classifiers (LR, DT, NB, SVM, RF, XGBoost, AdaBoost). However, the use of the univariate method may lead to neglecting the interaction between the features. The same correlation-coefficient analysis method for discovery of relationships between the features was used by Ganie et al. [18] for exploration of the PIMA data and achieved ~92.85% accuracy using gradient boosting. More comprehensive comparison study of filter-based methods (chi-square, Fisher's Score, Information Gain) and wrapper-based approaches (RF/XGBoost/GB/SVM/LR) has been carried out by Kaliappan et al. [3] on four datasets related to PIMA data.

Wrapper and embedded methods. Recursive Feature Elimination (RFE) appears recurrently as a solid baseline in numerous investigations. According to Li et al. [8], RFE was compared with the Genetic Algorithm on PIMA and the Iraqi Society Diabetes datasets; the latter appeared less efficient and less accurate than RFE, while the accuracy of RF decreased even further with the use of GA. Sahid et al. [13] implemented combinations of RF, LR, Mutual Information, PCA, ANOVA, and FDR using filter, wrapper, and embedded approaches for an Iraqi multi-class problem; the highest level of accuracy (96.4%) was gained with the use of SVM using filter approach on top-4 features – but the

were largely variable for the original and rebalanced versions of their dataset. Talari et al. [14] used information gain and correlation-based selection in combination with Bagging Decision Trees model.

Metaheuristic and hybrid feature selection. Bio-inspired optimization for feature selection is increasingly common in the literature, where higher accuracy is noted compared to the conventional filters. Rahman et al. [1] compared RFE, Grey Wolf Optimizer (GWO), Particle Swarm Optimization (PSO), Genetic Algorithm, and Boruta using PIMA and DiaHealth datasets and found Boruta to be the most effective approach (85.16% using LightGBM); importantly, the authors themselves have warned that this was because of the small sample size (768 cases and 8 features) of PIDD dataset and lack of external validation for two datasets used. Sirmayanti et al. [2] have used GWO with an autophagy mechanism (GWO-AM), achieving 90.91% accuracy in one PIMA example.

Ziane et al. [7] proposed the hybrid PSO-GWO technique with collaborative switching on three different datasets (PIMA, Frankfurt Hospital, Iraqi) with an optimal rate of 84.74%, indicating the dependency of their

on the demographics of the population in question that is not captured by single dataset analysis. K. et al. [9] obtained the best result for accuracy among all techniques (96.7%), which used PSO with a stacking ensemble, but due to the lack of information about the paper's authorship, data origin, and methodology, the reliability of these

cannot be confirmed.

Information-theoretic and mutual-information-based selection. Several studies favor mutual information as a redundancy-aware criterion. Tasin et al. [12] used MI-based selection with semi-supervised insulin imputation on a private Bangladeshi RTML dataset, pairing XGBoost with ADASYN oversampling (~81% accuracy) — a private, non-reproducible dataset that limits external verification. Waqas Khan et al. [15] combined chi-square, mutual information, and sequential feature selection across PIMA and an early-risk dataset, reporting an unusually high 99.35% accuracy with ANN; the absence of an external, unseen third dataset leaves overfitting or leakage difficult to rule out. Sukson et al. [10] took a union-based approach — combining LR significance, ReliefF, and MRMR — validated on both the UCI Early Stage Diabetes dataset and external hospital data; the drop from 97.69% internal to 85% external accuracy is presented by the authors as direct evidence of limited cross-population generalizability.

Ensemble-oriented feature selection with dimensionality reduction. Ngo et al. [16] combined Chi-Square, LASSO, Mutual Information, PCA, and RFE on a newly integrated U.S. diabetes dataset (deliberately avoiding reliance on PIMA alone), with MLP-plus-PCA performing best (93.07%); as a newly constructed multi-source dataset, however, it lacks a multi-year external benchmarking history. Sampath et al. [4] employed correlation-based feature selection with SMOTE and a hybrid of AdaBoost and XGBoost (accuracy 90.4%, AUC 0.968) and explicitly pointed out their dependence on structured tabular data as a limitation for any future unstructured data work. Zhou et al. [11] implemented the Boruta algorithm on the PIMA and Sylhet early risk data sets using a stacked classifier and achieved a very high 98% on PIMA, dataset.

Ansari et al. [17] examined various supervised learning algorithms to predict diabetes mellitus from Pima Indian Diabetes Dataset. The research conducted a comparative analysis of four supervised learning algorithms, namely SVM, RF, KNN, and NB, through 10-fold cross-validation and feature selection methods. In terms of prediction accuracy, SVM was found to be the most efficient classifier, followed by RF, KNN, and NB, thus proving the efficiency of supervised learning methods in predicting diabetes.

3. PROPOSED MULTISTAGE ENSEMBLE FEATURE SELECTION FRAMEWORK

The Multi-stage Ensemble Feature Selection (MEFS) technique employs three kinds of feature ranking methods, such as Nonlinear Dynamic Conditional Relevance Feature Selection (NDCRFS) [20], Hellinger Distance (HD) [22], and Feature Hamming Diversity (FHD)[21], to increase feature relevance, feature class discrimination, and avoid feature redundancy at once. The incorporation of the fusion approach in the proposed methodology presents various statistical views of feature subset selection for diabetes prediction

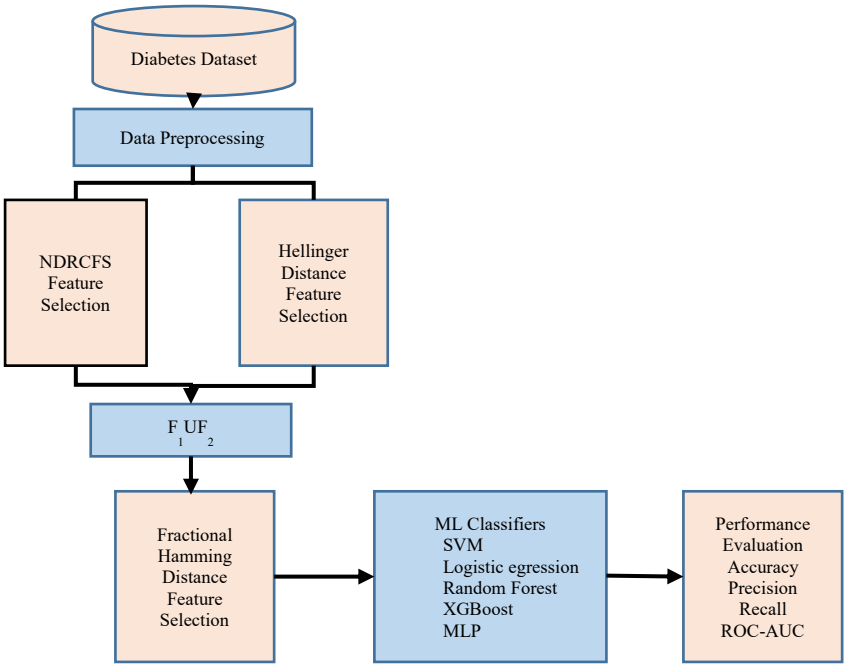

Figure 1. Proposed Multistage Ensemble Feature Selection Framework

Figure 1 illustrates the workflow of the proposed framework. The dataset is preprocessed to impute missing values, handle outliers and perform normalization. Next, the dataset is processed through the NDCRFS and Hellinger Distance algorithms individually, to select the two optimal feature subsets. These subsets are merged to form a single set of features, and then the fractional hamming distance algorithm is applied to eliminate any redundant features. Finally, the optimal feature subset is used to train multiple machine learning classifiers, and their performance is evaluated using standard classification metrics.

3.1 Stage I: Nonlinear Dynamic Conditional Relevance Feature Selection (NDCRFS)

Stage 1 involves the identification of informative features while also reducing redundancy of selected features. Filter approaches tend to assess the relevance of features independently and fail to consider the interaction of candidate features. Redundant features may, therefore, be ranked highly although they do not add much new information.

In order to solve this problem, the proposed framework utilizes the Nonlinear Dynamic Conditional Relevance Feature Selection (NDCRFS) [20]. It implements feature selection by means of greedy forward feature selection through mutual and conditional mutual information. In each iteration of this process, all candidate features get a relevance score determined by

$$\text{Score}(f_k) = MI(f_k, y) - \max_{f_s \in S}\left[CU(s,k)\left(CMI(f_k, f_s|y) - MI(f_s, y)\right)\right] \quad (1)$$

where

- f_k denotes the candidate feature,
- y represents the target,
- S is the previously selected feature subset,
- $MI(\cdot,\cdot)$ denotes Mutual Information,
- $CMI(\cdot,\cdot\,|y)$ represents Conditional Mutual Information.

The normalization coefficient is calculated as

$$CU(s,k) = \frac{2}{H(f_s)+H(f_k)+\varepsilon} \quad (2)$$

where

- $H(f)$ denotes the entropy of feature f,
- ε is a small positive constant to avoid divide by zero.

Before entropy calculation, continuous attributes undergo discretization using the process of uniform binning. The first component in Equation (1) is used for measuring dependency between an attribute and target class, while the second component imposes a penalty on the redundant information between the selected attributes. Thus, NDCRFS

in a ranked feature subset.

3.2 Stage II: Hellinger Distance-Based Feature Ranking

NDCRFS is successful in identifying informative features but it does not quantify the discriminative power of individual features. Thus, the next step determines the separability of features by using the Hellinger Distance (HD).

The feature probabilities for the respective classes are calculated from frequency distribution histograms. The Hellinger distance between diabetic and non-diabetic classes is calculated as

$$HD(P,Q) = \sqrt{\frac{1}{2}\sum_{i=1}^{n}\left(\sqrt{P_i} - \sqrt{Q_i}\right)^2} \quad (3)$$

where

- P_i denotes the probability of the i^{th} histogram bin associated with diabetic patients,
- Q_i denotes the probability of the i^{th} histogram bin associated with non-diabetic patients,
- n represents the total number of histogram bins.

Larger values of Hellinger Distance show that there is greater divergence between two class distribution and hence better discrimination capability. Ranking of features is done using their Hellinger distance score in decreasing order, and features having high rank form the second set of features. Hellinger distance is more robust to class imbalance than traditional statistical metrics for ranking of features since it measures the probability distribution of samples rather than counting of samples. It is well-suited to medical datasets where diseased classes are less frequent.

3.3 Stage III: Feature Hamming Diversity (FHD)

The first two phases mainly assess feature relevance and class discrimination. But, even features selected independently by these approaches may have information overlap. In order to increase diversity of selected features, the FHD approach is applied in the third phase.

First, all the continuous features are converted into discrete form using quantile binning. Then, the feature with maximum variance is selected as the representative feature. After that, each subsequent feature is selected based on its average Hamming distance from the already selected features.

The diversity score is calculated as

$$Score(f_c) = \frac{1}{|D|}\sum_{f_s \in D} H\,ammingDist(f_c, f_s) \quad (4)$$

where

- f_c denotes the candidate feature,
- D represents the set of already selected features,
- $HammingDist(\cdot,\cdot)$ computes the pairwise Hamming distance.

The normalized Hamming distance between two discretized feature vectors is computed as

$$HammingDist(f_i, f_j) = \frac{1}{N}\sum_{k=1}^{N} I\left(f_{ik} \neq f_{jk}\right) \quad (5)$$

where

- N is total samples,
- $I(\cdot)$ denotes the indicator function that returns one when the corresponding feature values differ and zero otherwise.

Features that have larger Hamming distance values provide complementary information and thus avoid redundancy in the selected features. Therefore, in this step, we obtain an ordering that focuses on diversity rather than class relevance.

3.4 Stage IV: Ensemble Feature Fusion

Each one of the three feature selection techniques separately produces a ranking of the relevant features. Because each one of these methods represents a distinct way of determining feature relevance, the

from these three techniques combined will yield a more reliable feature set.

Let

- R_1 denote the NDCRFS ranking,
- R_2 denote the Hellinger Distance ranking,
- R_3 denote the Feature Hamming Diversity ranking.

The voting score for each feature is computed as

$$Votes(f_i) = \sum_{m=1}^{3} I\left(f_i \in R_m\right) \quad (6)$$

where

$$I(f_i \in R_m) = \begin{cases} 1, & \text{if } f_i \text{ appears in ranking } R_m \\ 0, & \text{otherwise} \end{cases}$$

When two or more features obtain identical vote counts, the average rank is calculated as

$$AvgRank(f_i) = \frac{1}{Votes(f_i)}\sum_{m=1}^{3} Rank_m(f_i) \quad (7)$$

where $Rank_m(f_i)$ denotes the rank assigned by the m^{th} feature selection method.

Finally, the features are sorted according to the following priority:

1. Higher vote count.
2. Lower average rank.

The top-k ranked features constitute the final optimal feature subset used for classifier training.

Computational Complexity Analysis

For the data set has N samples with d features, the time complexity of NDCRFS algorithm is $O(d^2N)$ due to repeated computations of mutual information and conditional mutual information during the greedy feature selection process. In the case of Hellinger Distance stage, histograms need to be estimated for every feature, and the complexity of this stage is $O(dN)$. Feature Hamming Diversity calculates the Hamming distances of all pairs of discretized features, making the complexity as $O(d^2N)$. The final ensemble voting stage only requires ranking and aggregation operations with complexity$O(d\log d)$.

Despite the three-tier ranking process, the computational efficiency of the proposed framework is still feasible for moderately sized medical datasets because the classification algorithm trains the final model on a significantly smaller feature set. Also, the decrease in the dimensionality of features leads to lower memory requirements and classifier training times. This is demonstrated in the empirical study conducted on the Pima Indians Diabetes dataset where the proposed framework decreased the number of features from eight to five, while lowering memory requirements and training time.

4. EXPERIMENTAL SETUP

This section discusses the experiment design used to assess the efficiency of the Multistage Ensemble Feature Selection (MEFS) proposed for the purpose of diabetes prediction. The proposed MEFS methodology has been assessed using the PIMA Indian dataset [23 using several machine learning classifiers along with standard performance metrics and statistical significance tests.

4.1 Dataset Description

Experiments were performed on the Pima Indians Diabetes Dataset (PIDD) which is one of the most popular benchmark dataset used for testing the effectiveness of different diabetes prediction algorithms. Dataset is comprised of diagnostic features of female patients of Pima Indian descent with an age of 21 or above. Each instance contains eight input features and one output feature.

There are 768 records in total, which includes 500 records belonging to non-diabetic and 268 diabetic individuals, respectively. The clinical features include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age.

Table 1 summarizes the characteristics of the dataset.

Table 1. Characteristics of the Pima Indians Diabetes Dataset

Attribute	Description
Dataset	Pima Indians Diabetes Dataset
Number of Records	768
Number of Features	8
Target Classes	Diabetic / Non-Diabetic
Missing Values	Present in several clinical attributes
Feature Type	Numerical

4.2 Data Preprocessing

Missing values, varying scales of features, and class imbalance exist in medical data and have a negative impact on classifier accuracy. Thus, preprocessing was performed prior to feature selection.

At first, missing values for the Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI attributes were handled using the mean imputation method. The next step included performing the proposed Feature-Specific Mean Variance Normalization (FSMVN) [19], which normalizes each clinical attribute based on its distribution. In contrast to common normalization techniques, the Feature-Specific Mean Variance Normalization technique retains specific statistical information about the features' distributions, which is useful before the feature selection process.

As diabetic instances comprise a minority class, the Adaptive Synthetic Sampling (ADASYN) algorithm was used for balancing the distribution of classes. Adaptive Synthetic Sampling generates additional minority class samples around hard-to-classify decision boundaries, making it easier for classifiers to learn characteristics of the diabetic instances and avoiding class imbalance problems. The preprocessed dataset was then supplied to the proposed Multistage Ensemble Feature Selection framework.

4.3 Proposed Feature Selection Configuration

Each feature selection stage gives an individual rank list of informative features. The rank lists are then integrated via the voting-based technique presented in Section 3 in order to get the final set of features.

The proposed methodology is able to reduce the initial eight clinical attributes into five informative features with 37.5% reduction in the number of features and 37.5% reduction in the amount of memory used. The final set of selected features includes: Glucose, Pregnancies, Insulin, Diabetes Pedigree Function, and Age. These features always got high ranks from all the three feature selection techniques, which means their significant role in the prediction of diabetes.

4.4 Classification Models

The performance of the In order to analyze the performance and applicability of the suggested feature selection algorithm, five popular machine learning classifiers have been used.

Support Vector Machine (SVM)

Support Vector Machine determines the best separating hyperplane with maximum margin between diabetic and non-diabetic cases. Due to its ability to work well in higher dimensions, SVM is a good baseline classifier.

Logistic Regression (LR)

Logistic Regression calculates the probability of having diabetes by applying logistic function. Its simplicity, interpretability, and computational efficiency make it particularly suitable for clinical decision-support systems.

Random Forest (RF)

Random Forest is an ensemble classification model that contains several decision trees. The algorithm introduces randomness to features and uses bootstrapping for training.

Extreme Gradient Boosting (XGBoost)

XGBoost employs gradient boosting with regularization to construct sequential decision trees. It effectively models nonlinear relationships and has demonstrated excellent performance across numerous medical prediction tasks.

Multi-Layer Perceptron (MLP)

The Multi-Layer Perceptron is a feed forward neural network capable of learning complex nonlinear mappings between clinical variables and diabetes outcomes through back propagation. Using multiple classifiers enables an objective assessment of whether the proposed feature selection framework remains classifier-independent.

4.5 Model Validation Strategy

For obtaining correct and unbiased performance

, 10-fold cross validation was applied. Performance was calculated as an average of performance

of all the folds..

4.6 Performance Evaluation Metrics

The proposed framework was evaluated using several widely accepted classification metrics.

Accuracy

Accuracy represents the proportion of correctly classified samples.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

Precision

Precision measures the proportion of predicted diabetic patients who are actually diabetic.

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

Recall (Sensitivity)

Recall measures the ability of the classifier to correctly identify diabetic patients.

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

Specificity

Specificity measures the ability to correctly classify non-diabetic individuals.

$$Specificity = \frac{TN}{TN+FP} \quad (11)$$

F1-score

The F1-score balances Precision and Recall.

$$F1 = 2\left(\frac{Precision \times Recall}{Precision+Recall}\right) \quad (12)$$

ROC-AUC

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) measures the discrimination power of the classifier regardless of its threshold value. Apart from the prediction accuracy, the average training time was measured to determine the efficiency of computations after the feature selection.

4.7 Statistical Analysis

In order to investigate whether the statistical significance exists in the performance difference among various feature selection algorithms, non-parametric statistical tests have been used. Firstly, the Friedman Test has been carried out to analyze the average rank of different feature selection algorithms using several classifiers. The reason is that the performance of classifier does not always meet the requirements of the assumptions of the parametric test. Then, Wilcoxon Signed-Rank Tests with adjusted p-values have been employed for pairwise comparisons between the proposed framework and other feature selection algorithms.

4.8 Experimental Configuration

The overall experimental configuration adopted in this study is summarized in Table 2.

Table 2. Experimental Setup

Component	Configuration
Dataset	Pima Indians Diabetes Dataset
Preprocessing	Missing Value Mean Imputation, normalization FSMVN, ADASYN
Normalization	Feature-Specific Mean Variance Normalization (FSMVN)
Class Balancing	ADASYN
Feature Selection	NDCRFS + HD + FHD
Validation	10-Fold Cross Validation
Classifiers	SVM, LR, RF, XGBoost, MLP
Performance Metrics	Accuracy, Precision, Recall, Specificity, F1-score, ROC-AUC, Training Time
Statistical Tests	Friedman Test, Wilcoxon Signed-Rank Test

5. RESULT AND DISCUSSION

This section introduces the experimental study of the proposed Multistage Ensemble Feature Selection (MEFS) framework. The effectiveness of the proposed framework has been studied from the following four dimensions: (i) feature selection efficiency, (ii) classification efficiency by using various machine learning algorithms, (iii) performance comparison with current state-of-the-art feature selection methods, and (iv) statistical validation of the achieved

. All experiments have been conducted according to the experimental setup of Section 4 using 10-fold cross-validation.

5.1 Feature Selection Performance

The proposed MEFS model utilizes NDCRFS, Hellinger Distance, and Feature Hamming Diversity techniques to determine the most relevant clinical attributes for predicting diabetes disease. From the original eight attributes, the model determined five selected attributes, which maximize their relevancy, class separation, and diversity. Final selected features are glucose, pregnancies, insulin, diabetes pedigree function and age.

A reduction from eight to five attributes

in 37.5% reduction in feature dimensionality. The reduction further reduced memory consumption from 0.05 MB to 0.03 MB, accounting for 37% memory savings. This kind of feature dimensionality reduction is computationally efficient, and it ensures maintenance of relevant information needed to predict diabetes disease.

The NDCRFS stage focused mainly on the extraction of very important features in terms of their nonlinear dependencies with the output variable without causing any redundancy. Hellinger Distance focused on features that had high discrimination between the groups of diabetics and non-diabetics, whereas Feature Hamming Diversity filtered out those features that overlapped by choosing mutually different features.

5.2 Analysis of Individual Feature Ranking Methods

Tables 3 present the rankings generated independently by NDCRFS, Hellinger Distance, and Feature Hamming Diversity.

Table 3. Feature ranking

Selected Feature	NDCRFS Score	Hellinger Score	Fractional Hamming Distance Score
Glucose	0.12866	0.269497	0.373893
Age	-0.33918	0.277953	0.422066
Pregnancies	-0.37802	0.632998	0.445453
DiabetesPedigreeFunction	-0.41355	Not Selected	0.462166
Insulin	-0.42591	0.253921	0.444666
SkinThickness	Not selected	0.203291	Not selected

NDCRFS determined Glucose, Age, Pregnancies, Diabetes Pedigree Function, and Insulin as the most important features through the process of nonlinear conditional dependency. Hellinger Distance method chose Pregnancies, Age, Glucose, Insulin, and Skin Thickness as the most discriminative features due to their ability of separating classes. Hamming Diversity selected Glucose, Age, Insulin, Pregnancies, and Diabetes Pedigree Function as the most important features, which indicates that these features offer complementary not redundant information. Despite the fact that each feature selection technique utilizes its own criterion for evaluation, there are some common features that have been chosen as the most important in all three approaches, such as Glucose, Pregnancies, Age, and Insulin.

5.3 Classification Performance

The selected feature subset was evaluated using five widely adopted machine learning classifiers, namely Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multi-Layer Perceptron (MLP).

Table 4 compares classifier performance using the original feature set and the proposed ensemble-selected features.

Table 4. Performance comparison of machine learning classifiers using the complete feature set and the proposed MEFS-selected feature subset on the Pima Indians Diabetes Dataset.

Model	Accuracy		Precision		**Recall**		**Specificity**		**F1-Score**		**ROC AUC**		**Training_Time_seconds_avg**	
	All Features	Ensemble	All Features	Ensemble	All Features	Ensemble	All Features	Ensemble	All Features	Ensemble	All Features	Ensemble	All Features	Ensemble
SVM	0.7316	**0.7277**	0.5961	0.5799	0.7459	0.8285	0.724	0.674	0.6602	0.6799	0.825	0.8248	0.296	**0.115**
Logistic Regression	0.7524	**0.7342**	0.6334	0.599	0.7426	0.754	0.758	0.724	0.6788	0.6644	0.8376	0.8304	0.0205	**0.0128**
Random Forest	0.7629	**0.7407**	0.6477	0.6145	0.7158	0.7091	0.788	0.758	0.6776	0.6558	0.8208	0.8136	0.7061	**0.5819**
XGBoost	0.7264	**0.7251**	0.6004	0.603	0.6484	0.6491	0.768	0.766	0.6213	0.6233	0.7954	0.7955	0.1116	**0.0929**
MLP	0.746	**0.7329**	0.6155	0.5873	0.7423	0.8098	0.748	0.692	0.6706	0.6792	0.8344	0.8387	0.7255	**0.7184**

From the experiment

shows that there is no degradation of prediction quality by decreasing the dimension of the feature space. Instead, the suggested feature selection framework allows us to achieve prediction quality comparable to prediction quality by using all the features, but at the same time, we have decreased the computational complexity. Amongst the evaluated classifiers, Random Forest showed the best accuracy of classification with all the features, while XGBoost had stable

of prediction after feature selection with some insignificant change in the accuracy of prediction. At the same time, the MLP classifier showed better ROC-AUC after the feature selection, which suggests that the redundant features can negatively impact the work of neural networks. One important finding is the significant decrease in average training time of all classifiers after the feature selection. As fewer input features need to be considered during training, the computational complexity of the training process is reduced without compromising the ability to classify. This property is especially useful in resource-limited healthcare systems where there is a need for an efficient predictive model.

5.4 Comparative Analysis with Existing Feature Selection Methods

To further validate the effectiveness of the proposed framework, its performance was compared with several widely used feature selection techniques, including:

- Information Gain
- Chi-square
- ReliefF
- mRMR
- Boruta

- LASSO
- Random Forest Importance

The comparative evaluation considered prediction accuracy, F1-score, ROC-AUC, and average training time. Figure 2 shows the accuracy comparison with respect to various feature selection methods.

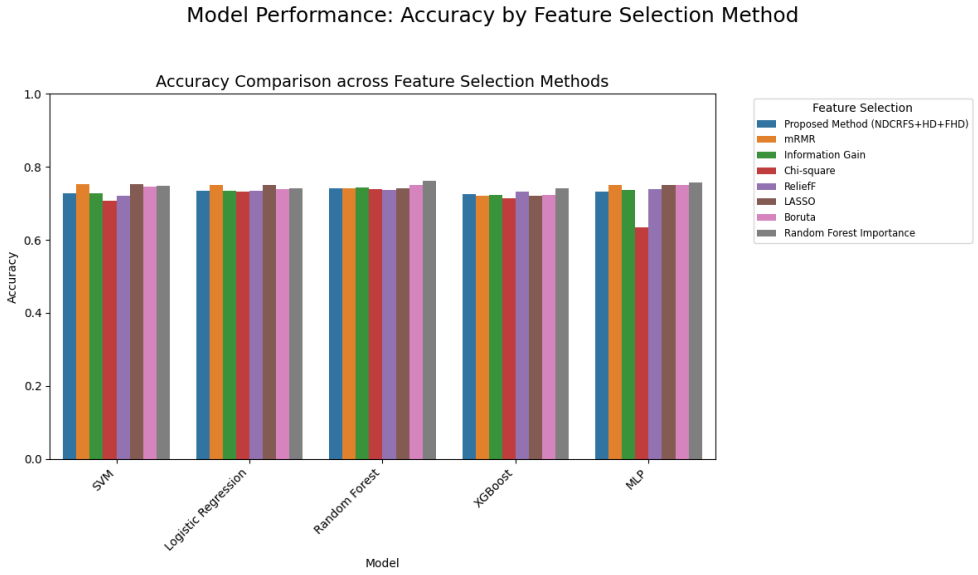

Figure 2. Reported performance of various feature selection methods and proposed MEFS framework.

The reveal that the developed MEFS method is able to achieve competitive predictive performance to other feature selection methods while identifying a smaller and more comprehensible set of features.

Figure 3 shows the training time comparison of the feature selection methods. While Random Forest Importance performed better in terms of classification accuracy, it consumed much longer training time and selected extra number of features. In contrast, the developed method is capable of making an effective balance between feature elimination and prediction since it considers three factors together. Unlike Information Gain and Chi-square, which focus on the relevance of individual features, the developed method offers enhanced redundancy elimination by applying NDCRFS and Feature Hamming Diversity. Likewise, the proposed method outperforms Relief in terms of ranking stability because of neighborhood selection sensitivity. This means that a combination of several feature assessment techniques can identify a more robust feature subset than using a single ranking technique.

5.5 Computational Efficiency Analysis

One of the key goals of feature selection is minimizing computational cost without affecting the accuracy of predictions. The proposed framework was able to meet the goal by reducing the initial feature space by 37.5%. It helped to achieve the following:

- Minimizing memory usage
- Decreasing classifier training time
- Minimizing the computational complexity
- Facilitating interpretability

Experimental

indicate that all the tested classifiers needed less training time post feature selection. This was especially true for the algorithms which are computationally expensive to use, such as Random Forest and Support Vector Machine. It clearly shows that the exclusion of unnecessary variables helps not only to develop a model but also to implement it in a clinical setting.

5.6 Statistical Validation

To analyze if the differences in performance obtained from the feature selection techniques are statistically significant, the Friedman test and Wilcoxon Signed-Rank test were applied. The result of the Friedman test showed a statistic value of 13.2034 with a p-value of 0.0673. As the p-value is greater than 0.05, the null hypothesis can be accepted, which means that the suggested feature selection technique has statistically equivalent performance compared with alternative approaches. In addition, the Wilcoxon Signed-Rank test was carried out for the pairwise comparison with adjusted p-values. The outcome of this analysis demonstrated that there are no statistically significant differences between the suggested MEFS approach and the other techniques for feature selection. Even though the suggested method selects a smaller number of features, its predictive performance remains the same as the performance of state-of-the-art methods.

5.7 Discussion

The of experiments show that the Multistage Ensemble Feature Selection model is able to solve many problems of the existing feature selection models. While the traditional feature selection filters select features using only one statistical measure, the proposed method considers feature relevance, separability, and diversity. As a result, redundant features are excluded, but important clinical information needed for an adequate prediction of diabetes is preserved. The selected set of features includes important clinical features like Glucose, Pregnancies, Age, Insulin, and Diabetes Pedigree Function. These features have been proven to be important factors of diabetes risk.

The other equally important achievement is the considerable decrease in computational cost. There is a significant reduction in feature dimensions, memory usage, and classifier training time without compromising on accuracy. This kind of computational efficiency proves to be very useful when it comes to real-time clinical decision support systems and in the future increment learning framework.

Even though this evaluation was done on the Pima Indians Diabetes Dataset, the presented model is independent and can easily be scaled up to large-scale multicenter clinical data. Moreover, the identified feature subset serves as a good basis for the next phase of this research, i.e., the incremental deep learning model and the XAI component.

6. CONCLUSION AND FUTURE WORK

In this work, MEFS is a multi-stage ensemble-based feature selection framework that combines Nonlinear Dynamic Conditional Relevance Feature Selection (NDCRFS), Hellinger Distance (HD), and Feature Hamming Diversity (FHD) via an ensemble voting scheme. In the experiment with the Pima Indians Diabetes dataset, MEFS successfully reduces the dimensionality of the original dataset by selecting only five out of eight (37.5%) features while retaining the predictive capabilities of MEFS with respect to Support Vector Machine, Logistic Regression, Random Forest, XGBoost, and Multi-Layer Perceptron classifiers with a decrease in computation cost and memory footprint. The comparative analysis of MEFS versus Information Gain, Chi-Square, ReliefF, mRMR, Boruta, LASSO, and Random Forest Feature Importance methods demonstrates that MEFS can achieve competitive predictive performance using fewer features. Moreover, the statistical testing based on Friedman and Wilcoxon Signed-Rank tests reveals that there is no statistically significant difference between the proposed framework and other feature selection methods. Therefore, MEFS is a reliable framework that selects relevant features for diabetes prediction with low complexity and model independence and can serve as a foundation for

The future work will concentrate on validation of MEFS through the use of large multi-center clinical data sets as well as broadening its application to different prediction problems. The future work will also involve investigating the potential use of XAI, incremental learning, and adaptive ensemble weights in the improvement of feature selection performance and stability.

REFERENCES

[1] Rahman, F., Hossain, S., Tiang, J.-J., & Nahid, A.-A. (2025). Diabetes prediction using feature selection algorithms and boosting-based machine learning classifiers. Diagnostics, 15(20), 2622. https://doi.org/10.3390/diagnostics15202622

[2] Sirmayanti, S., Prastyo, P. H., & Mahyati, M. (2025). Enhancing diabetes prediction performance using feature selection based on grey wolf optimizer with autophagy mechanism. Computer Methods and Programs in Biomedicine Update. https://doi.org/10.1016/j.cmpbup.2025.100207

[3] Kaliappan, J., Saravana Kumar, I. J., Sundaravelan, S., Anesh, T., Rithik, R. R., Singh, Y., Vera-Garcia, D. V., Himeur, Y., Mansoor, W., Atalla, S., & Srinivasan, K. (2024). Analyzing classification and feature selection strategies for diabetes

prediction across diverse diabetes datasets. Frontiers in Artificial Intelligence, 7, 1421751. https://doi.org/10.3389/frai.2024.1421751

[4] Sampath, P., Elangovan, G., Ravichandran, K., et al. (2024). Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique. Scientific Reports, 14, 28984. https://doi.org/10.1038/s41598-024-78519-8

[5] Asif, R., Upadhyay, D., Zaman, M., & Sampalli, S. (2025). Enhancing diabetes risk prediction: A comparative evaluation of bagging, boosting, and ensemble classifiers with SMOTE oversampling. Healthcare Analytics (ScienceDirect). https://www.sciencedirect.com/science/article/pii/S2352914825000498

[6] Giri, S. K., Dash, S., & Sahoo, T. (2024). An intelligent diabetes prediction system augmenting feature selection and balancing techniques. In Proceedings (Springer). https://doi.org/10.1007/978-981-99-5015-7_9

[7] Ziane, A., El Bouhissi, H., & Hanne, T. (2026). Intelligent diabetes prediction system based on hybrid PSO-GWO feature selection and optimized machine learning. Information, 17(6), 533. https://doi.org/10.3390/info17060533

[8] Li, X., Curiger, M., Dornberger, R., & Hanne, T. (2023). Optimized computational diabetes prediction with feature selection algorithms. In ISMSI 2023 (pp. 36-43). ACM. https://doi.org/10.1145/3596947.3596948

[9] Karuppasamy M, Jansi Rani M, Poorani K (2025),Metaheuristic Feature Selection for Diabetes Prediction with P-G-S Approach, Procedia Computer Science,Volume 252,2025,Pages 165-171,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2024.12.018.

[10] Sukson, P., Cholamjiak, W., Eiamniran, N., & Khwanmuang, M. (2026). Towards robust risk-based screening of early-stage diabetes: Machine learning models with union features selection and external validation. Diabetology, 7(1), 2. https://doi.org/10.3390/diabetology7010002

[11] Zhou, H., Xin, Y., & Li, S. (2023). A diabetes prediction model based on Boruta feature selection and ensemble learning. BMC Bioinformatics, 24, 224. https://doi.org/10.1186/s12859-023-05300-5

[12] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 10(1-2). https://doi.org/10.1049/htl2.12039

[13] Sahid, M. A., Babar, M. U. H., & Uddin, M. P. (2024). Predictive modeling of multi-class diabetes mellitus using machine learning and filtering Iraqi diabetes data dynamics. PLOS ONE, 19(5), e0300785. https://doi.org/10.1371/journal.pone.0300785

[14] Talari P, N B, Kaur G, Alshahrani H, Al Reshan MS, Sulaiman A, Shaikh A. Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2. PLoS One. 2024 Jan 18;19(1):e0292100. doi: 10.1371/journal.pone.0292100. PMID: 38236900; PMCID: PMC10796060.

[15] Waqas Khan, Q., Iqbal, K., Ahmad, R., Rizwan, A., Nawaz Khan, A., & Kim, D. (2024). An intelligent diabetes classification and perception framework based on ensemble and deep learning method. PeerJ Computer Science, 10, e1914. https://doi.org/10.7717/peerj-cs.1914

[16] Ngo, V. M., Bharot, N., Donohue, F., Kearney, P. M., Breslin, J. G., & Roantree, M. (2026). Analyzing and predicting diabetes with deep learning and visual insights. Intelligence-Based Medicine, 14, 100368. https://doi.org/10.1016/j.ibmed.2026.100368

[17] Ansari, G. A., Shafi, S., Ansari, M. D., & Shadab, A. (2025). Advanced supervised machine learning methods for precise diabetes mellitus prediction using feature selection. Frontiers in Medicine, 12, 1620268. https://doi.org/10.3389/fmed.2025.1620268

[18] Ganie, S. M., Pramanik, P. K. D., Malik, M. B., Mallik, S., & Qin, H. (2023). An ensemble learning approach for diabetes prediction using boosting techniques. Frontiers in Genetics, 14, 1252159. https://doi.org/10.3389/fgene.2023.1252159

[19] Skubleny, D., Ghosh, S., Spratlin, J., Schiller, D.E. and Rayat, G.R., "Feature-specific quantile normalization and feature-specific mean–variance normalization deliver robust bi-directional classification and feature selection performance between microarray and RNAseq data", BMC bioinformatics, vol. 25, no. 1, pp.136, 2024.

[20] Zhang, L., "A feature selection algorithm integrating maximum classification information and minimum interaction feature dependency information", Computational Intelligence and Neuroscience, no. 1, pp.3569632, 2021.

[21] Mahdian, M.A., Taheri, E., Rahbardar Mojaver, K. and Nikdast, M., "Hardware assurance with silicon photonic physical unclonable functions", Scientific Reports, vol. 14, no. 1, pp.25591, 2024.

[22] Shu, T., He, Z., Yin, X., Ding, Z. and Zhou, M., "Model-based diversity-driven learn-to-rank test case prioritization", Expert Systems with Applications, vol. 255, pp.124768, 2024.

[23] National Institute of Diabetes and Digestive and Kidney Diseases. (n.d.). *Pima Indians diabetes database* [Dataset]. Kaggle. Retrieved April 2025, from https://www.kaggle.com/datasets/uciml/PIMA-indians-diabetes-database