A Novel Approach for Diabetes Mellitus Detection Using a Modified Binary Multi-Neighbourhood Artificial Bee Colony Algorithm with Mahalanobis-Based Feature Selection (MBMNABC-Ma) and an Optimized Decision Forest Framework

Original Scientific Paper

Gaurav Pradhan *
Department of Computer Applications,
Sikkim Manipal Institute of Technology,
Sikkim Manipal University (SMU),
Majitar, India
gaurav.p@smit.smu.edu.in

Gopal Thapa
Department of Computer Applications,
Sikkim Manipal Institute of Technology,
Sikkim Manipal University (SMU),
Majitar, India
gopal.t@smit.smu.edu.in

*Corresponding author

Ratika Pradhan
Department of Computer Applications,
Sikkim University,
Gangtok , India
rpradhan01@cus.ac.in

Bidita Khandelwal
Department of General Medicine,
Sikkim Manipal Institute of Medical Sciences,
Sikkim Manipal University (SMU),
Tadong , India
bidita.k@smims.smu.edu.in

Abstract *– Diabetes is a critical global health issue caused by high blood sugar (hyperglycemia), leading to complications like cardiovascular disease, blindness, neuropathy, and kidney failure. Machine learning (ML) algorithms improve both the accuracy and efficiency of medical diagnoses. This study applies a Modified Binary Multi-Neighbourhood Artificial Bee Colony with Mahalanobis-based (MBMNABC-Ma) for a feature selection algorithm, combined with diverse ML models for diabetes identification. Compared to the conventional Binary Multi-Neighbourhood Artificial Bee Colony (BMNABC), MBMNABC-Ma improves classification accuracy and reduces computational complexity. Five diabetes datasets were analyzed using a 70-30% holdout cross-validation. The MBMNABC-Ma model, trained on Optimal Decision Forest (ODF) with Random Forest Ensemble (RFE), demonstrated high effectiveness. It achieved 97.23% accuracy on the Merged Datasets (comprising 130 US and PIMA datasets), 97.93% on the Iranian Ministry of Health Dataset, 96.05% on the Questionnaire Dataset, 98.39% on the Hospital of Sylhet Dataset, and 80.98% on the PIMA Dataset, with high specificity and sensitivity scores across all cases.*

Keywords: *Machine Learning, Feature Selection, Biomedical Data Analysis, Ensemble Learning, Diabetes Detection, Data Mining*

Received: July 8, 2025; Received in revised form: August 9, 2025; Accepted: August 19, 2025

1. INTRODUCTION

Diabetes is a common endocrine disease, defined by increased blood glucose levels caused by defects in the production or action of insulin or both [1]. The increase in diabetes cases has created a severe risk to the global healthcare system. An individual suffering from diabetes can have consequences like cardiovascular disease, blindness, renal failure, and neuropathy [2]. Hence, diabetes mellitus must be identified and treated as soon as possible to reduce the complications. In the meantime, prominent trends in diabetes-related early death were found through a study. Between 2000 and 2010, the death rate fell in high-income countries [3].

There was a further rise in death rates between 2010 to 2016 due to diabetes. In low-income countries, the rates of premature death continued to rise [4]. In India, over 77 million people suffer from diabetes, according to forecasts published in 2019 [5, 6]. Likewise, by 2045, there would be close to 134 million cases of diabetes in India, according to these estimates [7]. Diagnosing diabetes and its related diseases is important and cannot be exaggerated at an early age. Therefore, this offers a good opportunity for early intervention and better treatment tactics, which lowers the risk of serious issues, including heart disease and nerve damage [8]. Data Mining techniques and machine learning have become essential tools in detecting various diseases, specifically in the detection of diabetes. It aids with prediction, diagnosis, and complication management [9]. The strength and efficiency of machine learning come from its ability to find patterns that human experts might overlook, improving the diagnostic accuracy [10]. In order to create a successful ML model, feature selection helps in dimensionality reduction, model generalization, and computing efficiency [11, 12]. This study proposes a Modified Binary Multi-Neighbourhood Artificial Bee Colony algorithm with a Mahalanobis-based (MBMNABC-Ma) feature selection algorithm for the detection of diabetes. The study attempted to compare the performance against traditional algorithms by testing it on five diabetes datasets within an Optimal Decision Forest (ODF) framework. We evaluate the proposed model against other methods, including MBMNABC-Ma + k-NN, MBMNABC-Ma NB, MBMNABC-Ma + C4.5, MBMNABC-Ma + RS, and MBMNABC-Ma + SVM, using performance metrics like accuracy, specificity, and sensitivity. Results show that MBMNABC-Ma + ODF(RFE) outperforms most models in terms of accuracy and effectiveness.

1.1. LITERATURE REVIEW

Significant research work has been carried out on choosing features, imputation strategies, and managing values that are missing in the past; this section discusses and examines a few of the appropriate studies [10].

A. Negi et al. [11] used the UCI machine learning library to gather two of the datasets [13][14], those were then combined based on their similarities. Illustrations of unknown or missing data (unidentified data) were replaced with the value of 0. Some of the non-numeric entries were changed into numerical equivalents, and the features that were not relevant to the identification of diabetes were eliminated. A script that was included in LibSVM was employed to prepare the data, and the data points were normalized using a scale ranging from 0 to 1. The mean value was used to replace the amalgamated data points, and numerical values were assigned to symbolic representations. The split of 60%-40% of the dataset were considered for training and testing respectively. In the process of selecting relevant attributes, the Weka software platform was used to create the F-select script of the LibSVM package. A Support Vector Machine (SVM) was employed to create a predictive model. For validation, the training data were used through a 10-fold cross-validatio. Then, Feature selection was performed using both wrapper and ranker methods, resulting in 71% accuracy with the wrapper technique and 72% with the ranker technique. Moreover, the LibSVM F-select script was then employed, yielding 63% accuracy. Finally, by selecting all features, an accuracy of 72.92% was obtained. Heydari et al. [15] utilized a set of data consisting of 2536 occurrences from the University of Medical Sciences, Tabriz, to predict the presence of diabetes. Various methods, such as ANN, SVM, nearest neighbor, Bayesian network, and Decision Tree, were contrasted to identify the most effective procedure for diabetes diagnosis. Among all, the best accuracy was performed by ANN with 97.44%. On the other hand, 5-NN, Decision Tree, SVM, and Bayesian Network reached accuracy percentages of 81.19%, 95.03%, 90.85%, and 91.60%, respectively. Prerna et. al. [12] used online and offline surveys to generate a dataset containing eighteen inquiries about family history, lifestyle, and health. RStudio and the R programming language were implemented for analysis employing classifiers like k-Nearest Neighbor, Support Vector Machine, Decision Tree, Naïve Bayes, Logistic Regression, and Random Forest. The Random Forest produced the maximum accuracy of 94.10% and the dataset was split into 75% for the training and 25% for testing. Islam et al. [16] utilized machine learning methods, including Decision Trees, Naïve Bayes, Logistic Regression, and Random Forest for diabetes detection. 520 instances of the dataset were collected by direct investigations from Sylhet Diabetes Hospital. Using Cross-validation and an 80-20 split, Random Forest has achieved the finest accuracy with 97.4%. Dzulkalnine et al. [17] applied fuzzy principal factor analysis for feature selection from the PIMA dataset. In an 80-20 data split, the maximum accuracy of 72.078% was achieved using FPCA-SVM classification. A better hybrid imputation FPCA-SVM-FCM model was created, and accuracy measures revealed that FCM showed FCM surpassed SVM-FCM. Oladimeji et al. [18] obtained and used a set of data from Sylhet Diabetes Hospital, balancing with the SMOTE oversampling method. Abedini et al. [19] employed a PIMA dataset to propose an ensemble, hierarchical model. The models were individually trained and then integrated at an advanced level. At first, Decision Tree and Logistic Regression model were utilized, and next, feeding their results into a Neural Network for improved accuracy, an accuracy rate of 83% was obtained. Iyer et al. [20] uses Pima Indian Diabetes sets of data to forecast diabetes in women using Decision Tree (C4.5) and Naïve Bayes classifiers. Using the cross-validation and percentage split procedures, the data sets were divided into preparation and test sets. Moreover, 10-fold cross-validation was used to prepare the data with an achieved accuracy of 79.56%. Chang et al. [21] introduced a classification model suitable for elec-

tronic diagnostic systems. The three models: the C4.5 decision tree, Naïve Bayes, and Random Forest classifier were used on the diabetes datasets of Pima Indians. Position ranking, k-means clustering, and principal component analysis (PCA) were employed in the dataset analysis. Several matrices, such as accuracy, sensitivity, precision, F-score, specificity, and AUC (area under the curve), were used to evaluate the model's performance. On the entire dataset, Random Forest outperformed Naive Bayes and C4.5 decision trees, achieving 79.57% accuracy, 89.40% precision, 75.00% specificity, 85.17% f-score, and 86.24% AUC. Out of the three models, C4.5 achieved the highest sensitivity, at 88.43%.

Overall, research shows that ensemble-based classifiers and efficient feature selection methods are essential for accurate diabetes prediction. But the majority of current methods are either less accurate, have a lot of computing overhead, or aren't adequately generalizable to a variety of datasets. To improve the accuracy and efficiency of the detection of diabetes, Modified Binary Multi-Neighbourhood Artificial Bee Colony algorithm with Mahalanobis-based distance (MBMNABC-Ma), with Optimal Decision Forest and Random Forest Ensemble (ODF-RFE) has been proposed.

The paper is planned as follows: *Section 2* details the MBMNABC-Ma feature selection algorithm and the classification techniques used; *Section 3* presents the experimental results and comparisons with existing models; and *Section 4* concludes the study.

2. MATERIALS AND METHODOLOGY

This study aims to build an accurate and efficient system to detect diabetes by using advanced methods like model creation and feature selection. The approach uses the Random Forest Ensemble (RFE) along with the Optimized Decision Forest (ODF) for developing the model, and the MBMNABC-Ma algorithm is used to choose the most relevant features.

2.1. COLLECTION OF DATA

Various datasets were collected for validating the proposed model. The primary dataset, PIMA, and the secondary dataset, Diabetes 130-US hospitals, were provided by Negi et al.[11] from the UCI machine learning repository. These datasets span from 1999 to 2008 and were merged into one, with 102,536 participants—64,419 healthy, 38,115 unhealthy. The dataset's age range was 5 to 95 years, consisting of 47,055 males and 55,480 females. After filtering missing data, 5,000 instances were used. Another dataset from Heydari et al.[15] contained 2,209 individuals tested for type 2 diabetes in Tabriz, Iran, including 698 males and 1,837 females, ages 30 to 90, with 15 missing entries. A third dataset, provided by Neha Prerna et al. [12], included 952 participants aged 40-60, with 580 males and 372 females, of which 266 were diabetic and 685 non-diabetic. They also used the PIMA dataset. M. M. F. Islam et al.[16] introduced a dataset with 502 individuals (186 non-diabetic and 315 diabetic), aged 16 to 90, from the Sylhet Diabetes Hospital in Bangladesh. Finally, Kaggle [13] provided a dataset with 768 records, consisting of 500 non-diabetic and 268 diabetic cases, all female participants, aged 21 to 81. The study utilized binary classification datasets, each labeled with two classes: diabetic (1) and non-diabetic (0). These labels were originally included in the datasets and served as the ground truth during both training and evaluation. Table 1 summarizes these datasets.

Table 1. Datasets Used

Name of Dataset	Number of features	Instances	Classes
Merged Dataset (130 US and PIMA) [11]	46	5000	2
Iranian Ministry of Health [15]	19	2536	2
Questionnaire Dataset [12]	17	952	2
Hospital of Sylhet Dataset [16]	16	502	2
PIMA Dataset [13]	8	768	2

2.2. FEATURE SELECTION USING THE MODIFIED BINARY MULTI-NEIGHBORHOOD ARTIFICIAL BEE COLONY -MAHALANOBIS BASED (MBMNABC-MA) METHOD

The MBMNABC-Ma technique assessed the distance between neighbors *i* and *k* across all datasets using Mahalanobis distance rather than the traditional Euclidean or Hamming distance. The rationale behind adopting MBMNABC-Ma over existing feature selection techniques is its enhanced ability to identify optimal features that maximize classification accuracy while simultaneously minimizing the number of selected features. The classical BMNABC algorithm consists of the following stages: initialization of food sources for the bees; the employed bee phase, where each bee explores a new candidate solution based on neighborhood information; the onlooker bee phase, where bees select promising solutions based on the fitness of neighbors; and the scout bee phase, where new random solutions are introduced when stagnation is detected. In traditional BMNABC, the distance between the *ith* bee and its neighbors was calculated using Hamming distance, which may not adequately capture the relationships in datasets with continuous and correlated features. In contrast, the proposed MBMNABC-Ma algorithm employs Mahalanobis distance, which considers the covariance among features, providing a more reliable measure of dissimilarity between solutions. Moreover, the fitness evaluation process is improved by incorporating a multi-objective function that balances classification accuracy assessed through k-fold cross-validation and the number of selected features. Memetic Algorithms are hybrid optimization techniques that combine global search from algorithms like ABC with local refinement strategies to improve both convergence speed

and solution quality. In the MBMNABC-Ma method, the memetic part improves the top-performing solutions by locally adding or removing features to increase fitness. This hybrid structure helps the algorithm not just explore different regions of the search space globally, but also focus more closely on the best areas through local fine-tuning, resulting in better and smaller feature subsets. Through the use of Mahalanobis distance and a multi-objective fitness, the MBMNABC-Ma algorithm exhibited enhanced performance in both feature selection and classification accuracy, especially in the context of diabetes detection. The detailed working principles of MBMNABC-Ma using Mahalanobis distance measures and multi-objective optimization are outlined in Table 2 to Table 5.

Table 2. The Modified Binary Multi-Neighborhood Artificial Bee Colony with Mahalanobis (MBMNABC-Ma) Feature Selection using Mahalanobis Distance Measure

INPUT		
Dataset:	$X \in R^{N \times D}$	Diabetes mellitus dataset with N samples and D features
Class Labels:	$Y \in \{0,1\}^N$	Binary class labels, where 1 indicates a diabetic and 0 a non-diabetic
Parameters:	N_{pop}	Number of candidate solutions
	T	Maximum number of iterations
	R_{max}	Maximum neighbourhood radius for identifying far neighbours using Mahalanobis distance
	Cls	A supervised classifier for fitness evaluation
	k	Number of folds for cross-validation
	α, β	Weights for multi-objective fitness
OUTPUT		
A reduced subset of features $F_{best} \subseteq F$ that provides the highest classification accuracy and minimizes the number of features for detecting diabetes.		
PROCESS		
Step 1.	**Initialize the Population**	
	Step 1.1.	Generate initial solutions $x_1, x_2, \ldots, xN_{pop}$ where each solution x_i is a binary vector of length D (representing selected features): $x_{ij} = \begin{cases} 1 & if\ rand\ (0,1) \geq 0.5 \\ 0 & otherwise \end{cases}$ for $i = 1,2, \ldots, N_{pop}$ and $j = 1,2, \ldots \ldots, D$ (1)
	Step 1.2.	Evaluate the fitness $f(x_i)$ of each solution using multi-objective fitness: Classifier trained on the selected features corresponding to x_i. Perform k-fold cross-validation and compute the average accuracy $Acc(x_i)$ Compute fitness: $f(x_i) = \alpha \times Acc(x_i) - \beta \times \frac{Number\ of\ Selected\ Features\ in\ x_i}{D}$ (2)
Step 2.	**Far Neighbour Exploration and New Solution Generation**	
for t=1 ***to*** T ***do***		
	Step 2.1.	**Far Neighbour Identification** For each solution x_i compute Mahalanobis distance to every other solution x_k: $MD(x_i, x_k) = \sqrt{(x_i - x_k)^{-T} S^{-1} (x_i - x_k)}$ (3) where S is the covariance matrix of the feature data. Identify x_k as a far neighbor if: $MD(x_i, x_k) \geq max(MDi) \geq R \times mean(MD_i)$, where R is a neighborhood radius that is updated dynamically: $R = R_{max}\left(1 - \frac{t}{T}\right)$ (4)
	Step 2.2.	**Generate a New Candidate Solution** Call the Neighbor-Based Solution Generation Algorithm (Table 3) with the current solution x_i and its far neighbors x_k (identified using Mahalanobis distance) to generate a new solution v_i. $v_i = NeighborBasedSolution\ Generation\ (x_i, x_k)$ (5)
	Step 2.3.	**Selection Based on Fitness** Calculate the selection probability p_i for each solution x_i based on its fitness $p_i = \frac{f(x_i)}{\sum_{k=1}^{N_{pop}} f(x_k)}$ (6) //Use these probabilities to probabilistically select solutions for further exploration
	Step 2.4.	**Local Search on Top Solutions (Memetic Search)** Every M iteration, perform local refinement on top k solutions: Try adding or removing one feature at a time. Accept changes that improve the fitness $f(x)$.
	Step 2.5.	**Explore Near Neighbors** Call the Fitness-Based Neighbor Exploration Algorithm (Table 4) to explore nearby solutions using Mahalanobis distance for neighbor selection and generate a new solution v_i $v_i = FitnessBased\ Neight)$ (7)
	Step 2.6.	**Scout Bee Exploration and Solution Replacement**
	Step 2.6.1.	**Identify Stagnating Solutions** If a solution does not improve after a certain number of trials (stagnation), it is marked for replacement.
	Step 2.6.2.	**Generate a New Random Solution** Call the Random Solution Generation Algorithm (Table 5) to replace stagnating solutions with a new random solution x_i^{new}. $x_i^{new} = RandomSolution\ Generation\ (x_i)$ (8)
	Step 2.7.	**Memorize the Best Solution** After each iteration, keep track of the solution x_{best} with the highest accuracy.
end		

Step 3.	**Best feature subset F_{best} identification**
	Return $F_{best} \subseteq F$ the subset of features for the solution x_{best} for which the highest accuracy was received.

The neighbor-based solution generation of the MBMNABC-Ma feature selection approach is designed to explore the feature space effectively by leveraging the information from far neighbors identified using Mahalanobis distance. After identifying the far neighbors in Step 2.2 of the MBMNABC-Ma algorithm, the Neighbor-Based Solution Generation Algorithm is invoked to generate a new candidate solution by utilizing the structure of the best-performing neighbors. By considering the data's covariance structure through Mahalanobis distance, the algorithm ensures that neighbors are more meaningfully selected based on feature interdependencies. This leads to a more informed and efficient exploration of the search space, increasing the probability of discovering feature subsets that enhance overall fitness. The detailed process flow of neighbor-based solution generation has been outlined in Table 3.

Table 3. The Neighbor-Based Solution Generation Algorithm

INPUT	
• Current solution x_i	
• Set of far neighbors x_k, identified using Mahalanobis distance	
OUTPUT	
New candidate solution v_i	
PROCESS	
Step 1.	**Compute Average Best Neighbor**
	Compute the APB_{ij} of the far neighbors x_k $APB_{ij} = \frac{1}{N_{far}} \sum_{k=1}^{N_{far}} p^{best_{kj}}$ (9) //where $p^{best_{kj}}$ denotes the best solution found so far for the k^{th} far neighbor
Step 2.	**Generate New Solutions**
	Generate a new solution v_i using the best information from the far neighbors: $v_{ij}=x_{ij}+rand(0,1)\times(x_{ij}-APB_{ij})$ (10)
Step 3.	**Convert to Binary**
	Convert v_{ij} into a binary value for feature selection $v_{ij} = \begin{cases} 1 & if\ rand\ (0,1) \geq 0.5 \\ 0 & otherwise \end{cases}$ (11)
Step 4.	**Return New Solution**
	Return the new solution v_i to MBMNABC

Similar to the Neighbor-Based Solution Generation Algorithm, the Fitness-Based Neighbor Exploration Algorithm, as described in Table 4, also aims to enhance the feature space exploration. However, instead of focusing on far neighbors, it concentrates on near neighbors, which are solutions that are close to the current solution x_i based on Mahalanobis distance and have relatively high fitness values. This method refines the search by exploiting the best-performing nearby solutions.

Table 4. The Fitness-Based Neighbor Exploration Algorithm

INPUT	
• Current solution x_i	
• Set of far neighbors x_k, identified using Mahalanobis distance	
OUTPUT	
New candidate solution v_i	
PROCESS	
Step 1.	**Select Best Neighbor**
	Select the best neighbor x_{best_k} from the set of near neighbors based on fitness.
Step 2.	**Generate New Solutions**
	Generate a new solution v_i using the best near neighbor: $v_{ij}=x_{ij}+rand(0,1)\times(x_{ij}-x_{best_{kj}})$ (12)
Step 3.	**Compare and update fitness**
	Compare the fitness of the new solution $f(v_i)$ with the current solution $f(x_i)$ *if* $f(v_i)>f(x_i)$ then $x_i=v_i$ (13) *end*
Step 4.	**Return updated solution x_i**

In contrast to both the Neighbor-Based Solution Generation Algorithm and the Fitness-Based Neighbor Exploration Algorithm, which rely on the information from neighboring solutions, the Random Solution Generation Algorithm is used when a current solution has stagnated, i.e., it fails to improve after several trials. This algorithm seeks to provide variety to the search space and avoid it from getting stuck in local optima.

Table 5. The Random Solution Generation Algorithm

INPUT	
Current solution x_i marked for replacement	
OUTPUT	
New random solution x_i^{new}	
PROCESS	
Step 1.	**Generate Random Solution**
	For each feature *j*, generate a random binary value for the new solution: $x_{ij}^{new} = \begin{cases} 1 & if\ rand\ \ (0,1) \geq 0.5 \\ 0 & otherwise \end{cases}$ (14)
Step 2.	**Return the new random solution x_{ij}^{new}**

The MBMNABC-Ma algorithm for feature selection operates through multiple stages of exploration and exploitation, guided by neighborhood-based solution generation, fitness-based exploration, and random solution generation. Together, these methods ensure that the algorithm effectively navigates the feature space to identify an optimal subset of features that maximizes classification accuracy while minimizing redundancy. At the core of the MBMNABC-Ma algorithm the Mahalanobis distance measure is used to define

the proximity between solutions in the search space. Unlike hamming distance, the Mahalanobis distance adjusts distances according to the covariance structure of the data and considers feature correlations [22]. This provides a more accurate representation of feature relationships in continuous-valued datasets, enabling more precise identification of significant and non-redundant features. Consequently, MBMNABC-Ma is particularly effective in feature selection tasks involving real-world datasets where features exhibit interdependencies, such as in diabetes detection.

To evaluate the effectiveness of the MBMNABC-Ma algorithm, feature selection was independently performed on each of the five diabetes datasets. These datasets contain heterogeneous features ranging from medical test results and medication usage to lifestyle indicators. The algorithm was applied after preprocessing, enabling the identification of feature subsets most relevant to classification while eliminating redundant or irrelevant attributes. Table 6 presents the summary of selected features for each dataset, listing the total number of features, the count of those selected by the algorithm, and the specific feature names retained.

The results indicate that MBMNABC-Ma effectively adapts to varying dataset structures. For instance, the Merged Dataset (130-US + PIMA), originally containing 46 features, was reduced to 21 highly informative variables, such as Repaglinide, Pioglitazone, Change, Readmitted, Race, and Tolbutamide. Similarly, the Iranian Ministry of Health dataset retained 11 out of 19 features, with BMI, Triglyceride, and Cholesterol appearing as the most influential. In the Questionnaire dataset, 15 out of 17 features were preserved, including lifestyle-related variables like Family_diabetes, BMI, Stress, and Sleep. Notably, all 16 features from the Sylhet Hospital dataset were selected, highlighting the clinical relevance of symptoms such as Polydipsia, Polyuria, Alopecia, and visual blurring. Likewise, the PIMA dataset retained all 8 of its original features, suggesting their collective importance in diabetes detection.

This dataset-specific selection underscores the adaptability and robustness of the MBMNABC-Ma algorithm across varying feature spaces. It also demonstrates the algorithm's capacity to uncover both clinical and behavioral indicators that are strongly associated with diabetes, thereby enhancing model interpretability and predictive power.

Table 6. Dataset-wise Feature Selection Results Using the Modified Binary Multi-Neighbourhood Artificial Bee Colony with Mahalanobis-based (MBMNABC-Ma) Algorithm

Dataset	Total Features	No. of Selected Features	Features Name
Merged Dataset (130-US + PIMA) [11]	46	21	Repaglinide, Pioglitazone, Change, Readmitted, Race, Tolbutamide, Gender, Age, A1Cresult, DP_function, Weight, Max_glu_serum, Pregnancy, Troglitazone, Glipizide, Citoglipton, TriFold_Skin Thickness, Acetohexamide, Examide, BMI,
Iranian Ministry of Health [15]	19	11	BMI, Triglyceride, Cholesterol, Weight, HDL, History_of_pregnancy, FBS, Result_of_high_blood_pressure_screening, Age, History_of_diabetes, Family_history_of_diabetes
Questionnaire Dataset [12]	17	15	Family_diabetes, BMI, Age, Stress, Physically_active, Sleep, Soundsleep, Urinationfreq, Regularmedicine, Bplevel, Alcohol, Pregnancies, Gender, Highbp, Junkfood
Hospital of Sylhet Dataset [16]	16	16	Polydipsia, Polyuria, Age, Gender, Sudden_weight_loss, Irritability, Alopecia, Weakness, Itching, Polyphagia, Visual_blurring, Delayed_healing, Genital_thrush, Muscle_stiffness, Obesity, Partial_paresis
PIMA Dataset [13]	8	8	Glucose, Age, Insulin, Pregnancies, BloodPressure, BMI, SkinThickness, DiabetesPedigreeFunction

2.3. PROPOSED DIABETES DETECTION MODEL COMBINING MBMNABC-MA AND OPTIMIZED DECISION FOREST

The proposed method for diabetes detection uses a combination of two powerful techniques: Modified Binary Multi-Neighbourhood Artificial Bee Colony with Mahalanobis Distance (MBMNABC-Ma) for selecting features and an Optimized Decision Forest (ODF) for classification. The process begins by preparing the diabetes dataset through steps like dealing with missing values, normalization, and encoding. MBMNABC-Ma identifies the most important features by using Mahalanobis distance, which takes into account the relationships between features and the dataset's internal structure. This makes it highly effective for continuous datasets where feature interdependence plays a major role in prediction accuracy. The algorithm starts with an initial group of feature subsets and evaluates them based on how well they support classification. It explores both nearby and distant features in the dataset to ensure a thorough search and maintains diversity by introducing new random solutions when progress slows down. This cycle continues until the best feature subset, F_{best} is found.

Once the relevant features are selected, the ODF classifier is applied for predicting diabetes. ODF is an improved version of the Random Forest algorithm. It enhances performance by selecting only the most important decision trees to form a subforest, which helps reduce computation time and improve accuracy. Since ODF focuses on key features identified by MBMNABC-Ma, it provides better results, minimizes redundancy, and improves the model's ability to generalize. Additionally, it makes the prediction process more transparent, helping medical professionals understand how decisions are made, a valuable benefit in healthcare applications.

Overall, combining MBMNABC-Ma for feature selection with ODF for classification results in a highly accurate and efficient system for detecting diabetes. It handles high-dimensional medical data effectively, offers strong prediction performance in terms of accuracy, sensitivity, and specificity, and maintains clarity in decision-making, all of which are essential in clinical settings. The proposed model is illustrated in Fig. 1.

3. RESULTS AND DISCUSSION

This research compares feature selection using the conventional BMNABC algorithm and the Modified BMNABC with Mahalanobis distance (MBMNABC-Ma) across five transcontinental diabetes datasets. All the datasets were first preprocessed using the KNN imputation method and then passed through the Optimized Decision Forest (ODF) framework with the help of the Random Forest Ensemble (RFE) algorithm.

The performance of the proposed MBMNABC-Ma + ODF (RFE) method was compared with other classifiers like MBMNABC-Ma + k-Nearest Neighbors (kNN), MBMNABC-Ma + Support Vector Machine (SVM), MBMNABC-Ma + Naïve Bayes, MBMNABC-Ma + Rough Set (RS), and MBMNABC-Ma + C4.5 decision tree. Comparative results were also analyzed against the conventional BMNABC and previously published research. Accuracy, specificity, and sensitivity were employed as evaluation metrics. The MBMNABC-Ma + ODF (RFE) approach demonstrated superior performance across all metrics and datasets, highlighting its potential for robust, real-world diabetes detection applications. To ensure a strong comparison, key performance metrics were analyzed, including the Receiver Operating Characteristic (ROC) curve. These illustrations make it easier to distinguish the models' capabilities in various contexts. To comprehensively examine the diagnostic capabilities of each model, a detailed evaluation of key metrics, namely accuracy, sensitivity, and specificity, was performed. This rigorous comparison offers a clear understanding of the advantages and distinctive strengths of the proposed MBMNABC-Ma, combined with ODF using RFE about competing approaches. An educated viewpoint on the suggested model's possible real-world applications is facilitated by this thorough assessment, which provides insightful information on the model's efficacy and dependability in the context of diabetes diagnosis.

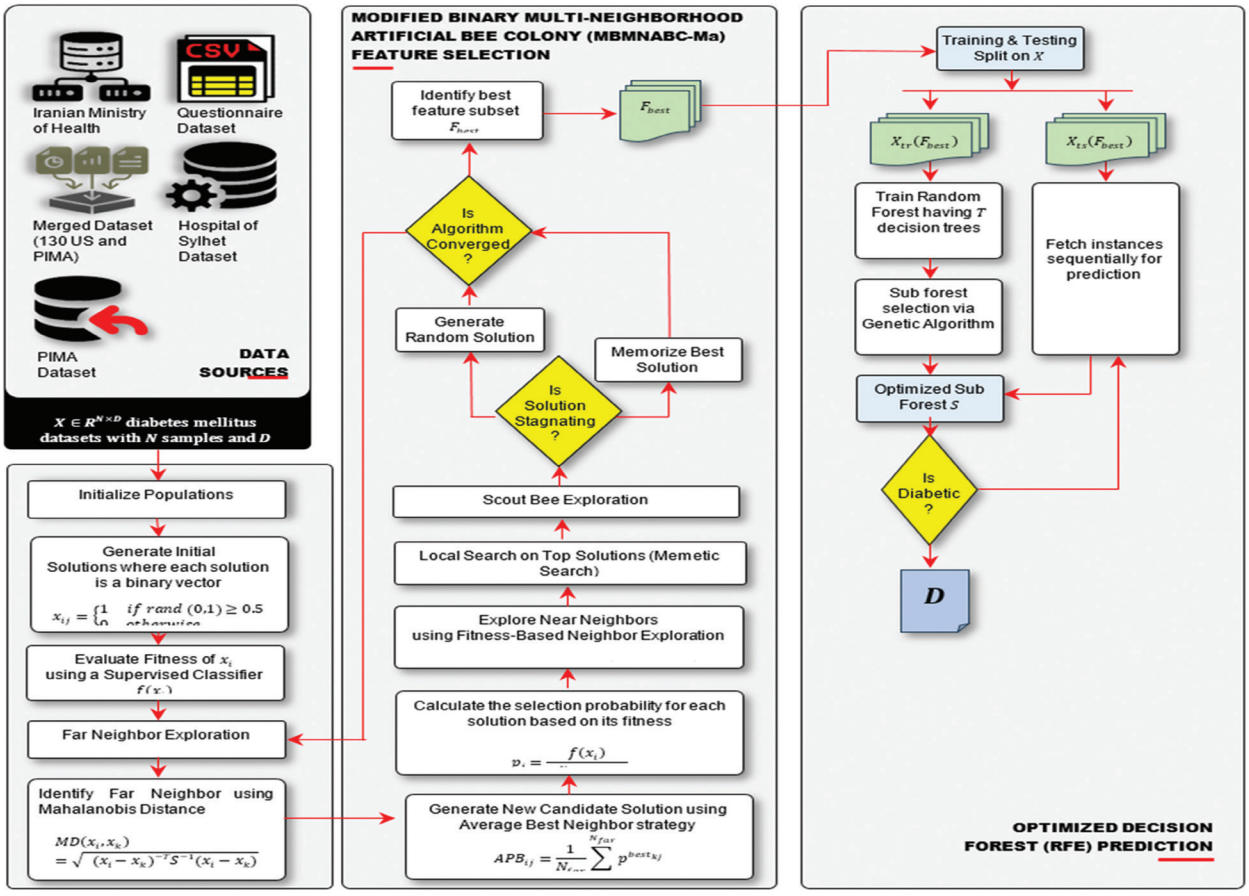

Fig. 1. Block diagram of the proposed diabetes detection model

A thorough analysis of several models, including MBMNABC-Ma + C4.5, MBMNABC-Ma + k-NN, MBMNABC-Ma + NB, MBMNABC-Ma + RS, and MBMNABC-Ma + SVM, in combination with the MBMNABC-Ma + ODF (RFE) model, is shown in Fig. 2. The combined dataset, which incorporates information from both the US and PIMA sources, is used for this evaluation. A useful illustration of the performance evaluation performed on the

Iranian Ministry of Health dataset may be seen in Fig. 3. The assessment findings from the questionnaire dataset are further displayed in Fig. 4, emphasizing the models' comparative capabilities. Additionally, the results of the performance evaluation carried out on the Sylhet Diabetes Hospital dataset are presented in Fig 5.

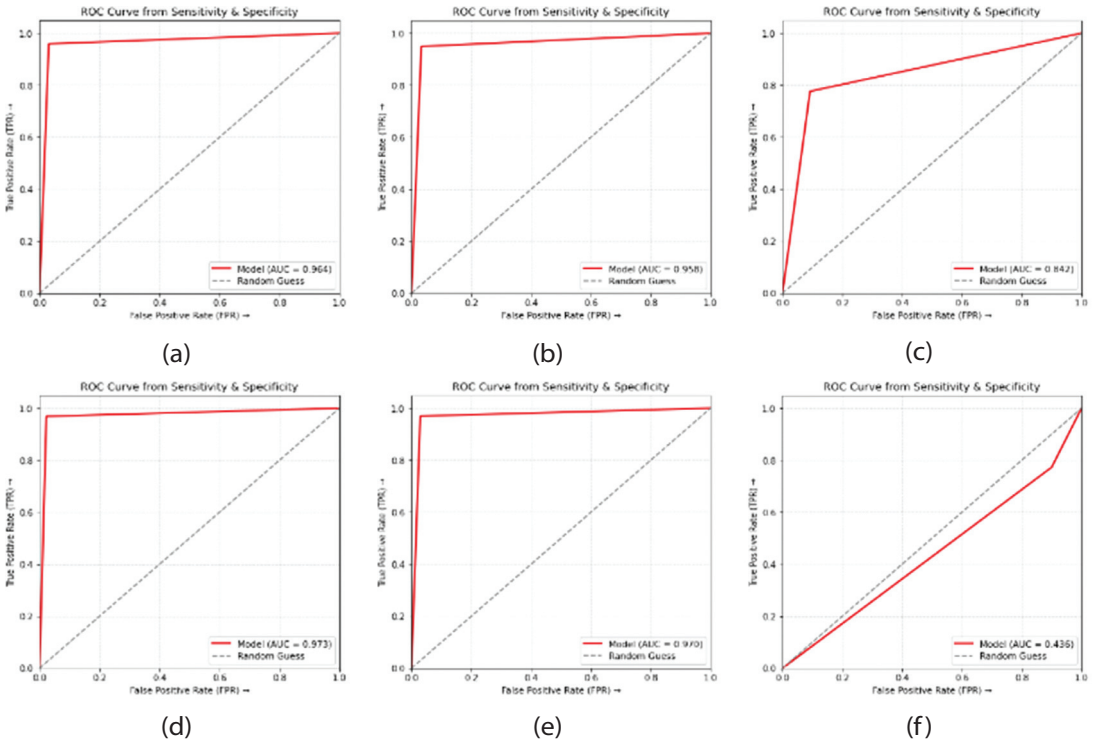

Fig 2. Comparative performance analysis of the models (a) MBMNABC-Ma + C4.5, (b) MBMNABC-Ma + k-NN, (c) MBMNABC-Ma + NB, (d) MBMNABC-Ma + ODF(RFE), (e) MBMNABC-Ma + RS, and (f) MBMNABC-Ma + SVM evaluated on the merged dataset (130 US and PIMA samples)

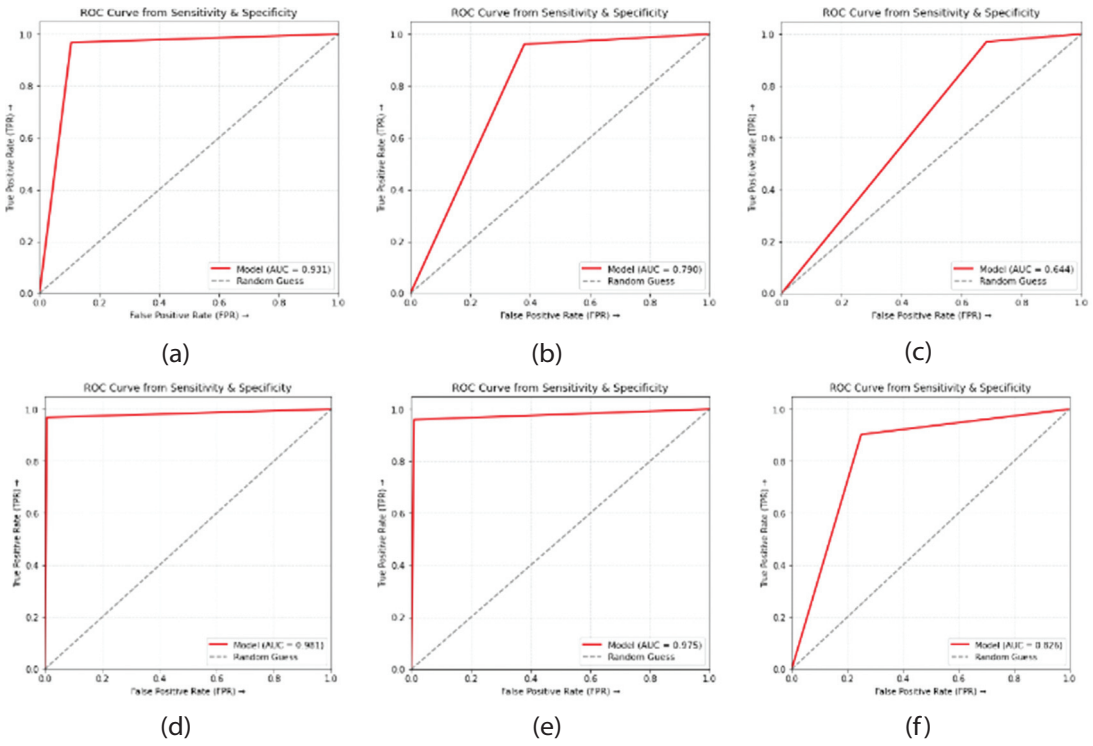

Fig 3. Comparative performance analysis of the models (a) MBMNABC-Ma + C4.5, (b) MBMNABC-Ma + k-NN, (c) MBMNABC-Ma + NB, (d) MBMNABC-Ma + ODF(RFE), (e) MBMNABC-Ma + RS, and (f) MBMNABC-Ma + SVM evaluated on the Iranian Ministry of Health dataset

Lastly, Fig 6 presents a comprehensive visualization of the performance evaluation conducted on the PIMA dataset, offering a clear and comparative perspective on the effectiveness of the models across different datasets.

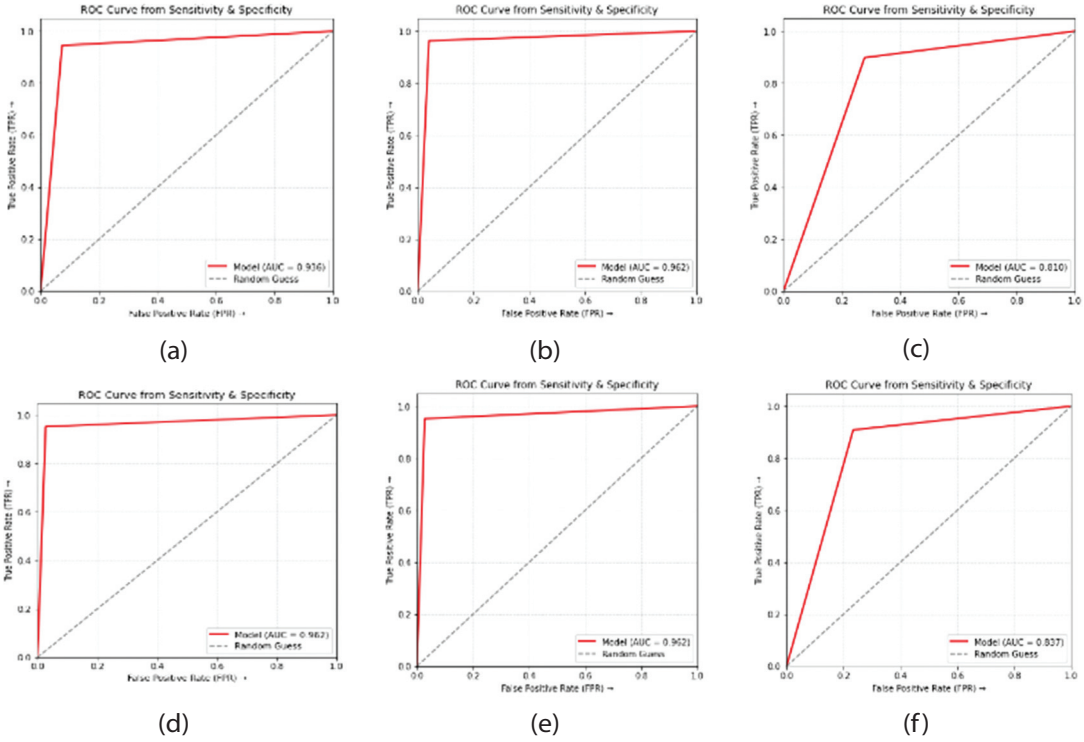

Fig 4. Comparative performance analysis of the models (a) MBMNABC-Ma + C4.5, (b) MBMNABC-Ma + k-NN, (c) MBMNABC-Ma + NB, (d) MBMNABC-Ma + ODF(RFE), (e) MBMNABC-Ma + RS, and (f) MBMNABC-Ma + SVM evaluated on the Questionnaire Dataset

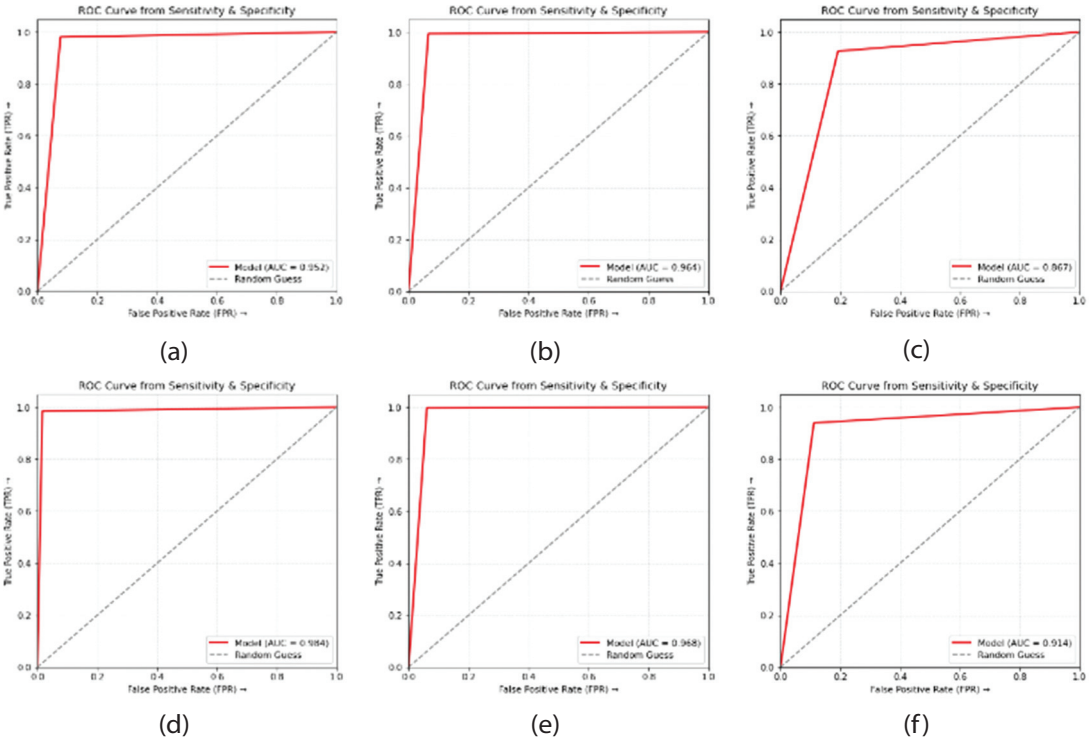

Fig 5. Comparative performance analysis of the models (a) MBMNABC-Ma + C4.5, (b) MBMNABC-Ma + k-NN, (c) MBMNABC-Ma + NB, (d) MBMNABC-Ma + ODF(RFE), (e) MBMNABC-Ma + RS, and (f) MBMNABC-Ma + SVM evaluated on the Sylhet Diabetes Hospital of Sylhet Dataset

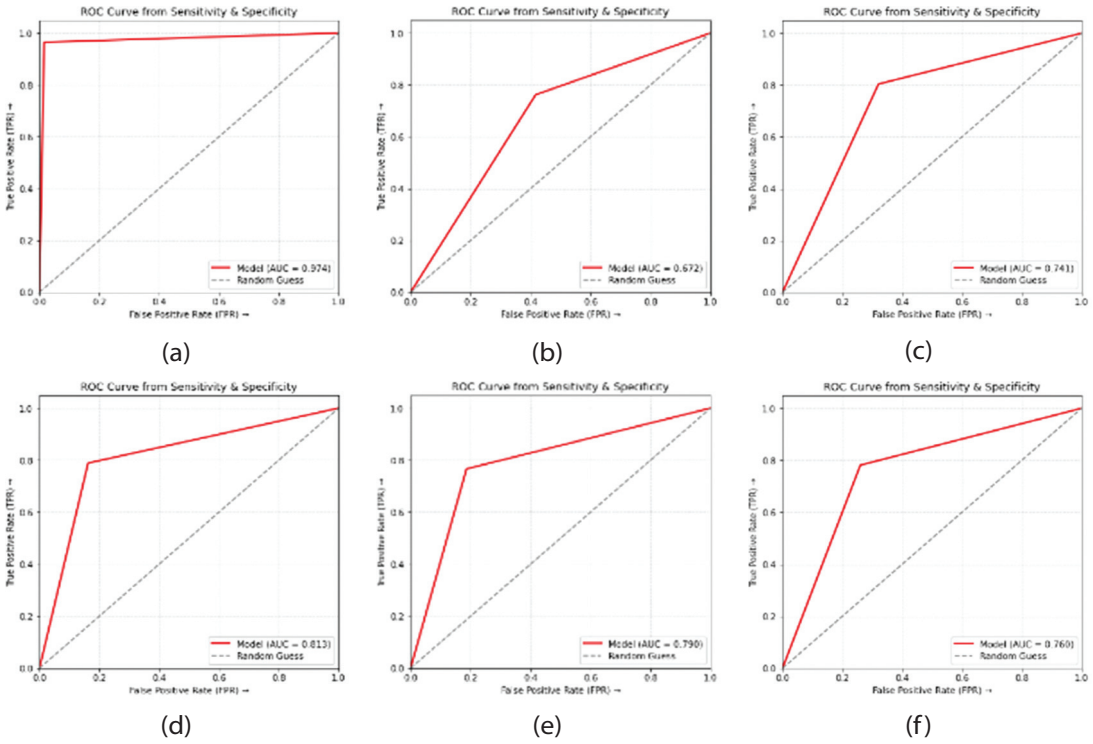

Fig 6. Comparative performance analysis of the models (a) MBMNABC-Ma + C4.5, (b)MBMNABC-Ma + k-NN, (c) MBMNABC-Ma + NB, (d) MBMNABC-Ma + ODF(RFE), (e) MBMNABC-Ma + RS, and (f) MBMNABC-Ma + SVM evaluated on the PIMA dataset

The above figures comprehends the subtle differences in performance of each model across various datasets, which helps to provide a thorough grasp of their prospective applications and probable ramifications in the field of diabetes diagnosis.

3.1. PERFORMANCE OUTCOMES WITH VARIOUS METHODS

The comparative analysis evaluated models based on accuracy, specificity, and sensitivity. The models assessed include MBMNABC-Ma + Random Forest (RF), MBMNABC-Ma + k-Nearest Neighbors (k-NN), MBMNABC-Ma + Naïve Bayes (NB), MBMNABC-Ma + C4.5, MBMNABC-Ma + Rough Set (RS), and MBMNABC-Ma + Optimized Decision Forest (ODF) using Random Forest Ensemble (RFE). The results for the Merged Dataset (130 US and PIMA records) are presented in Table 7, while the performance outcomes for the Iranian Ministry of Health dataset, the Questionnaire Dataset, the Hospital of Sylhet Dataset, and the PIMA dataset are detailed in Tables 8, 9, 10, and 11, respectively.

In Table 7, the MBMNABC-Ma + ODF (RFE) algorithm achieved the highest accuracy (97.23%) for diabetes detection, outperforming all other compared methods. MBMNABC-Ma + Naïve Bayes achieved an accuracy of 82.06%, MBMNABC-Ma + SVM achieved 83.94%, MBMNABC-Ma + k-NN reached 95.68%, MBMNABC-Ma + C4.5 attained 96.37%, and MBMNABC-Ma + Rough Set (RS) scored 96.98%. For specificity, MBMNABC-Ma + SVM achieved the highest value at 100%, followed by MBMNABC-Ma + ODF (RFE) at 97.75%, MBMNABC-Ma + C4.5 at 96.95%, and MBMNABC-Ma + RS at 97.06%. High specificity indicates the ability of the model to accurately identify healthy individuals, thereby reducing false positives. In terms of sensitivity, MBMNABC-Ma + RS (96.93%) and MBMNABC-Ma + ODF (RFE) (96.82%) performed the best, followed by MBMNABC-Ma + C4.5 (95.91%) and MBMNABC-Ma + k-NN (94.89%). Meanwhile, MBMNABC-Ma + SVM (77.27%) and MBMNABC-Ma + Naïve Bayes (77.64%) exhibited the lowest sensitivity, suggesting their potential challenges in accurately detecting all diabetes cases. High sensitivity is crucial to ensure diabetic individuals are correctly identified, minimizing the occurrence of false negatives.

Table 7. Performance of Proposed Detection Methods on the Merged Dataset (130 US and PIMA records) (10-Fold Cross Validation)

Detection Method	Accuracy (%)	Specificity (%)	Sensitivity (%)
MBMNABC-Ma + C4.5	96.37	96.95	95.91
MBMNABC-Ma + kNN	95.68	96.7	94.89
MBMNABC-Ma + NB	82.06	90.78	77.64
MBMNABC-Ma + ODF(RFE)	**97.23**	97.75	96.82
MBMNABC-Ma + RS	96.98	97.06	96.93
MBMNABC-Ma + SVM	83.94	100.00	77.27

Table 8. Performance of Proposed Detection Methods on the Iranian Ministry of Health dataset (10-Fold Cross Validation)

Detection Method	Accuracy (%)	Specificity (%)	Sensitivity (%)
MBMNABC-Ma + C4.5	96.19	89.41	96.76
MBMNABC-Ma + kNN	92.57	61.95	96.07
MBMNABC-Ma + NB	81.34	31.7	97.02
MBMNABC-Ma + ODF(RFE)	**97.93**	99.29	96.82
MBMNABC-Ma + RS	97.37	99.11	95.97
MBMNABC-Ma + SVM	90.22	75	90.25

The MBMNABC Ma combined with ODF using RFE achieved the highest accuracy of 97.93 percent, specificity of 99.29 percent, and sensitivity of 96.82 percent on the Iranian Ministry of Health dataset, demonstrating its strong capability for making precise predictions, as presented in Table 8. The MBMNABC-Ma + RS method closely shadowed by achieving 97.37% accuracy, with the specificity of 99.11% and 95.97% sensitivity. Comparing other methods, such as MBMNABC-Ma + k-NN and MBMNABC-Ma + Naïve Bayes, showed lesser accuracy and specificity, with MBMNABC-Ma + Naïve Bayes particularly displaying a very low specificity (31.7%). Hence, the result emphasizes that the MBMNABC-Ma + ODF(REF) and MBMNABC-Ma + RS performed extremely well in this dataset, showing significant potential for the precise and reliable detection of diabetes on the Iranian Ministry of Health dataset.

Table 9. Performance of Proposed Detection Methods on the Questionary dataset (10-Fold Cross Validation)

Detection Method	Accuracy (%)	Specificity (%)	Sensitivity (%)
MBMNABC-Ma + C4.5	94.01	92.65	94.48
MBMNABC-Ma + kNN	**96.21**	96	96.3
MBMNABC-Ma + NB	84.76	72.16	89.84
MBMNABC-Ma + ODF(RFE)	96.05	97.27	95.18
MBMNABC-Ma + RS	96.05	97.27	95.18
MBMNABC-Ma + SVM	86.86	76.6	90.83

Table 9, with the Questionary dataset, specifies significant insights into the efficiency of various methods for diabetes detection. In this dataset, the MBMNABC-Ma + kNN method gave the admirable accuracy of 96.21%, with strong specificity (96%) and sensitivity (96.3%). The methods, MBMNABC-Ma + ODF(REF) and MBMNABC-Ma + RS, gave the same results in terms of accuracy (96.05%), specificity (97.27%), and sensitivity (95.18%). The MBMNABC-Ma + C4.5 also gave a good accuracy of 94.01%, with specificity (92.65%) and sensitivity (94.48%). The other method, MBMNABC-Ma + NB and MBMNABC-Ma + SVM, provided lower accuracy, specificity, and sensitivity. This highlights that the MBMNABC-Ma + ODF(REF) and MBMNABC-Ma + RS performed best in the detection of diabetes in the Questionary dataset.

Table 10. Performance of Proposed Detection Methods on the Hospital of Sylhet dataset (10-Fold Cross Validation)

Detection Method	Accuracy (%)	Specificity (%)	Sensitivity (%)
MBMNABC-Ma + C4.5	95.8	92.31	98.04
MBMNABC-Ma + kNN	97	93.4	99.34
MBMNABC-Ma + NB	87.82	80.79	92.62
MBMNABC-Ma + ODF(RFE)	**98.39**	98.39	98.41
MBMNABC-Ma + RS	96.36	93.92	99.66
MBMNABC-Ma + SVM	92.01	88.83	93.93

Using the Hospital of Sylhet dataset, Table 10 highlights the performance of various algorithms for diabetes detection. In this dataset, the MBMNABC-Ma + ODF(REF) methods gave the highest accuracy (98.39%) as compared to other methods, with the specificity and sensitivity of 98.39% and 98.41% respectively. Similarly, the MBMNABC-Ma + kNN achieved an accuracy of 97% with a specificity of 93.4% and a sensitivity of 99.39%. Although the MBMNABC-Ma + RS did not provide the best accuracy and specificity as compared to MBMNABC-Ma + ODF(REF) and MBMNABC-Ma + kNN but it achieved the highest sensitivity of 99.66%. The MBMNABC-Ma + NB method gave the lowest accuracy, specificity, and sensitivity of 87.82%, 80.70% and 92.62% respectively, demonstrating that MBMNABC-Ma + ODF(REF) and MBMNABC-Ma + kNN emerged as the top methodologies for diabetes detection in the Hospital of Sylhet dataset.

In Table 11, using the PIMA dataset, the MBMNABC-Ma + ODF (RFE) method achieved the highest accuracy of 80.98% specificity of 83.74% and sensitivity of 78.88% indicating that it is efficient in accurately detecting diabetes in this dataset, The MBMNABC-Ma + RS achieved the second highest accuracy with 78.66%, reasonable specificity of 81.35% and notable sensitivity of 76.65%. The MBMNABC-Ma + c4.5 method gave the

highest sensitivity of 96.44% and specificity of 98.45% but notably lower accuracy of 76.17%. Hence, these results suggest that MBMNABC-Ma + ODF (RFE) and MBMNABC-Ma + RS performed best in the detection of diabetes in the PIMA dataset.

Table 11. Performance of Proposed Detection Methods on the PIMA Dataset (10-Fold Cross Validation)

Detection Method	Accuracy (%)	Specificity (%)	Sensitivity (%)
MBMNABC-Ma + C4.5	76.17	98.45	96.44
MBMNABC-Ma + kNN	70.44	58.30	76.20
MBMNABC-Ma + NB	76.43	67.90	80.38
MBMNABC-Ma + ODF(RFE)	**80.98**	83.74	78.88
MBMNABC-Ma + RS	78.66	81.35	76.65
MBMNABC-Ma + SVM	77.08	73.96	78.13

3.2. COMPARATIVE ANALYSIS WITH EXISTING TECHNIQUES

In Table 12, the proposed MBMNABC-Ma + ODF(RFE) method achieves a remarkable accuracy of 97.23%. The proposed methodology considerably gave better results as compared to the conventional BMNABC + ODF(RFE) reported by Pradhan et al. [23] on the Merged Dataset (130 US and PIMA), which had an accuracy of 96.36%. The substantial improvement of 0.87% emphasizes the effectiveness of MBMNABC-Ma in refining feature selection processes. Other methods, like SMOTE + Random Forest by Pradhan et al.[24] with the accuracy of 84.60% and LIBSVM by Negi et al. [11] with an accuracy of 73.00%, further illustrates the strength of the proposed method. This enhancement not only reinforces its potential as a leading method in the field but also highlights its ability to deliver superior classification outcomes compared to the traditional approach.

Table 12. Comparative result analysis between Conventional BMNABC and MBMNABC-Ma for the Merged Dataset (130 US and PIMA)

Dataset	Authors	Methods	Accuracy (%)
Merged Dataset (130 US and PIMA) [11]	Negi et al. [11]	SVM (Classification) + LIBSVM (Feature Selection)	73.00
	Pradhan et al. [23]	BMNABC + ODF(RFE)	96.36
	Pradhan et al. [24]	Random Forest + SMOTE	84.60
	Proposed	MBMNABC-Ma + ODF(RFE)	**97.23**

In Table 13, the proposed method MBMNABC-Ma + ODF(RFE) performed better than the BMNABC + ODF(RFE) by Pradhan et al. [23], which stated an accuracy of 97.93% accuracy in the Iranian Ministry of Health Dataset. Even though the proposed method performs better than several advanced methods, it's important to identify that MBMNABC-Ma's benefits are obtained from its flexibility and creative feature selection skills. For instance, Heydari et al. [15] achieved an accuracy of 97.44% with expert feature selection combined with ANN, Pradhan et al. [24] with SMOTE combined with Random Forest, achieved 96.80% accuracy, and Habibi et al. [25] reached 97.60% using expert feature selection with C4.5. The little discrepancy in accuracy shows that the proposed approach not only beats the competition but also provides substantial value in terms of the interpretability and pertinence of characteristics. The little discrepancy in accuracy shows that the proposed approach not only beats the competition but also provides substantial value in terms of the interpretability and pertinence of characteristics.

Table 13. Comparative result analysis between Conventional BMNABC and MBMNABC-Ma for the Iranian Ministry of Health

Dataset	Authors	Methods	Accuracy (%)
Iranian Ministry of Health [15]	Heydari et al. [15]	ANN + Expert Feature selection (Manual)	97.44
	Habibi et al.[25]	C4.5 + Expert Feature selection (Manual)	97.60
	Pradhan et al. [24]	Random Forest + SMOTE	96.80
	Pradhan et al.[23]	BMNABC + ODF(RFE)	97.28
	Proposed	MBMNABC-Ma + ODF(RFE)	**97.93**

The MBMNABC-Ma + kNN methodology acquires an accuracy of 96.21% for the Questionnaire Dataset in Table 14, falling only 0.22% short of the top-performing technique, BMNABC + ODF(RFE) by Pradhan et al. [23], which recorded 96.43%.

Table 14. Comparative result analysis between Conventional BMNABC and MBMNABC-Ma for the Questionnaire Dataset

Dataset	Authors	Methods	Accuracy (%)
Questionnaire Dataset [12]	Tigga et al. [12]	Random Forest	94.10
	Pradhan et al. [24]	Random Forest + SMOTE	93.70
	Pradhan et al. [23]	BMNABC + ODF(RFE)	96.43
	Proposed	MBMNABC-Ma+ kNN	**96.21**

However, it does not reach this benchmark, the proposed technique still displays improved performance compared to the BMNABC + ODF (RFE), which obtains 96.43% accuracy. In comparison, Tigga et al [12] achieved 94.10% accuracy with Random Forest and Pradhan et al. [24] reported 93.70% with SMOTE combined with Random Forest. This tiny difference exhibits how competitive MBMNABC-Ma is at achieving appropriate characteristics and classifying data.

The close performance emphasizes MBMNABC-Ma's possibility for additional evolution and offers positive options for improvement.

Table 15 shows the impressive accuracy of 98.39% achieved by the MBMNABC-Ma + ODF(RFE) in the Hospital of Sylhet Dataset. This is still not as accurate as the greatest recorded accuracy of 99.23% by Gündoğdu et al. [26], but it is significantly more accurate than the BMNABC + kNN technique of 97.01% by Pradhan et al. [23]. This improvement of 1.38% underlines the efficacy of the proposed strategy in addressing the challenges of this dataset. Other methodologies, like SMOTE with Random Forest by Pradhan et al.[24] which achieved an accuracy of 98.10% and SFS + ANN by Buyrukoğlu et al. [27] with an accuracy of 99.10% further emphasis the strength of the proposed method. This improvement of 1.38% underlines the efficacy of the proposed strategy in addressing the challenges of this dataset. MBMNABC-Ma's dynamic performance demonstrates how it may improve feature selection and emphasizes its usefulness in intricate real-word situations.

Table 15. Comparative result analysis between Conventional BMNABC and MBMNABC-Ma for the Hospital of Sylhet Dataset

Dataset	Authors	Methods	Accuracy (%)
Hospital of Sylhet Dataset [16]	Islam et al. [16]	Random Forest	97.4
	Pradhan et al. [24]	Random Forest + SMOTE	98.10
	Buyrukoğlu et al. [27]	ANN + SFS	99.10
	Nipa et al. [28]	APGWO-based MLP	97.00
	Gündoğdu et al. [26]	XGBoost + Random forest	99.23
	Prasanth [29]	XG Boost + SelectKBest	98.00
	Yasar [30]	FFNN + CSA	99.04
	Ma [31]	Neural Network + Min–Max	96.20
	Saboor et al. [32]	Optimize Selection + kNN + SMOTE	93.66
	Elsadek et al.[33]	Random Forest + Supervised Attribute Filter	97.88
	Rony et al. [34]	Random Forest + CFS	97.50
	Hasan et al. [35]	Extra Trees + PCC	99.06
	Pradhan et al. [23]	BMNABC + kNN	97.01
	Proposed	MBMNABC-Ma + ODF(RFE)	**98.39**

Finally, in Table 16, the proposed methodology (MBMNABC-Ma + ODF(RFE) achieved an accuracy of 80.98% which is more as compared to the BMNABC + ODF(RFE) result of 77.21% published by Pradhan et al. [23]. For instance, methodologies employing mean imputation and Naïve Bayes reported by Mousa et al. [36] with an accuracy of 85.00% and Chang et al. [21] with an accuracy of 79.13% it yields lower accuracies. In the PIMA dataset, an existing technique varying from 72.90% to 79.13%, the MBMNABC-Ma + ODF(RFE) method shows a competitive advantage. The results show that the proposed methodology, MBMNABC-Ma, is flexible and robust in different datasets: the Merged Dataset (130 US and PIMA), the Iranian Ministry of Health, the Questionnaire dataset, the Hospital of Sylhet, and the PIMA datasets. Although the proposed methodology falls short of the existing benchmarks, its strength lies in its flexibility, capability, and robustness in detecting diabetes.

Table 16. Comparative result analysis between Conventional BMNABC and MBMNABC-Ma for the PIMA Dataset

Dataset	Authors	Methods	Accuracy (%)
PIMA Dataset [13]	Mousa et al. [36]	LSTM + Mean imputation	85.00
	Rajni et al. [37]	RB-Bayes + Mean imputation	72.90
	Chang et al. [21]	Naïve Bayes + PCA	79.13
	Sisodia et al.[38]	Naïve Bayes + Mean imputation	76.30
	Pradhan et al. [23]	BMNABC + ODF(RFE)	77.21
	Proposed	MBMNABC-Ma + ODF(RFE)	**80.98**

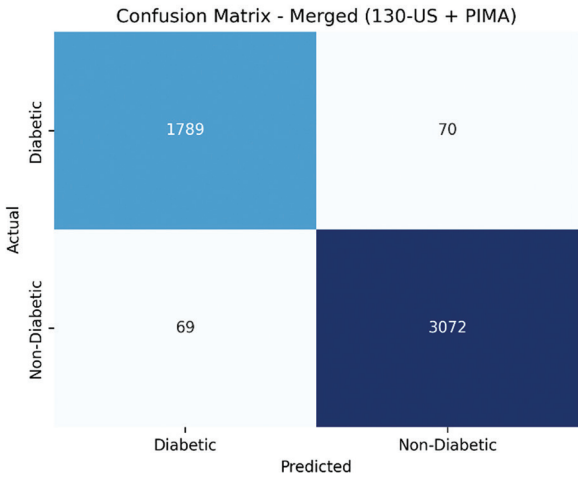

Fig 7. Confusion matrix for the Merged Dataset showing the classification performance of the proposed MBMNABC-Ma+ ODF(RFE) model

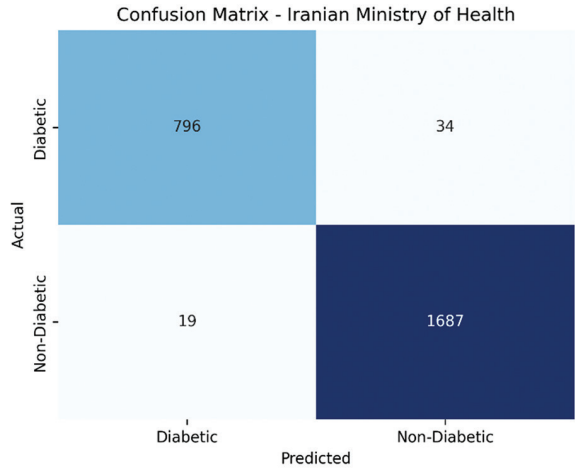

Fig 8. Confusion matrix for the Iranian Ministry of Health Dataset showing the classification performance of the proposed MBMNABC-Ma+ ODF(RFE) model

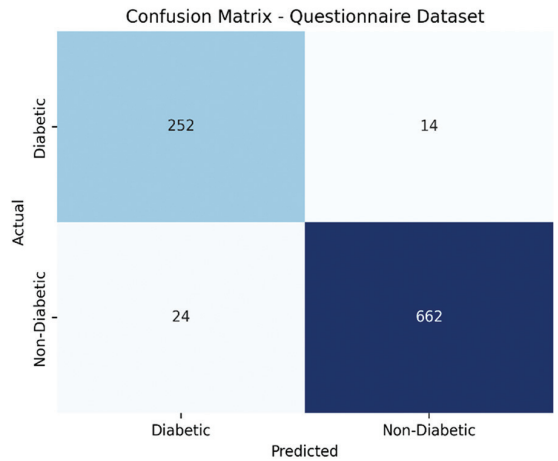

Fig 9. Confusion matrix model for the Questionnaire Dataset showing the classification performance of the proposed MBMNABC-Ma + ODF(RFE) model

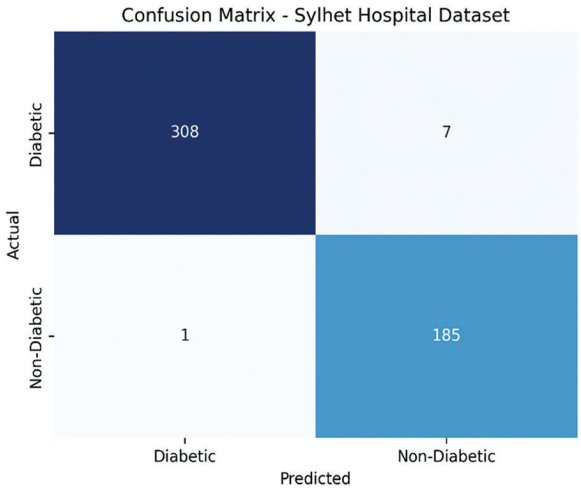

Fig 10. Confusion matrix for the Sylhet Diabetes Hospital Dataset showing the classification performance of the proposed MBMNABC-Ma + ODF(RFE) model

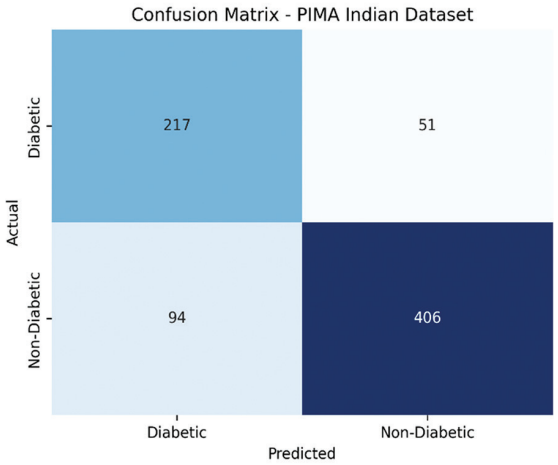

Fig 11. Confusion matrix model for the PIMA Dataset showing the classification performance of the proposed MBMNABC-Ma + ODF(RFE) model

4. CONCLUSION

Diabetes remains a significant public health concern, particularly among adults and elderly individuals, where early detection plays a vital role in reducing the risk of severe complications. This study explored the effectiveness of a novel meta-heuristic feature selection approach for diabetes detection by leveraging five diverse datasets containing a rich set of clinical and demographic variables. Through comprehensive experimentation, the proposed MBMNABC-Ma combined with the Optimized Decision Forest (RFE) framework demonstrated superior performance in terms of accuracy, sensitivity, and specificity compared to traditional methods. These findings not only confirm the robustness of the proposed approach but also underscore its potential for practical implementation in real-world clinical settings.

Moreover, this research emphasizes the importance of advanced feature selection techniques in improving

model precision and reducing redundancy. Looking ahead, future work may incorporate explainability techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to better understand the predictions made by complex models like ODF, thereby enhancing interpretability and trust in healthcare applications. Additionally, integrating deep learning architectures such as Convolutional Neural Networks (CNNs) may further refine detection capabilities by capturing deeper and more abstract patterns in high-dimensional medical data.

5. REFERENCE

[1] A. Ramachandran, C. Snehalatha, A. Raghavan, A. Nanditha, "Classification and Diagnosis of Diabetes", Textbook of Diabetes, Wiley, 2024, pp. 22-27.

[2] Md. A. Uddin et al. "Machine Learning Based Diabetes Detection Model for False Negative Reduction", Biomedical Materials & Devices, Vol. 2, No. 1, 2024, pp. 427-443.

[3] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, F. H. Juwono, "Diabetes Detection Based on Machine Learning and Deep Learning Approaches", Multimedia Tools and Applications, Vol. 83, No. 8, 2023, pp. 24153-24185.

[4] L. Al Rayes, M. Haggag, I. Afyouni, "Predicting Pre-Diabetic and Diabetes in Adults and Elderlies Using Machine Learning", Proceedings of Advances in Science and Engineering Technology International Conferences, Abu Dhabi, UAE, 3-5 June 2024, pp. 1-8.

[5] S. Luhar et al. "Lifetime risk of diabetes in metropolitan cities in India", Diabetologia, Vol. 64, No. 3, 2021, pp. 521-529.

[6] G. Pradhan, R. Pradhan, B. Khandelwal, "A Study on Various Machine Learning Algorithms Used for Prediction of Diabetes Mellitus", Proceeding of the International Conference on Computing and Communication, 2020, pp. 553-561.

[7] R. R. A. Bourne et al. "Causes of vision loss worldwide, 1990-2010: A systematic analysis", Lancet Glob Health, Vol. 1, No. 6, 2013, pp. e339-e349.

[8] M. T. Moghaddam et al. "Predicting Diabetes in Adults: Identifying Important Features in Unbalanced Data Over a 5-year Cohort Study Using Machine Learning Algorithm", BMC Medical Research Methodology, Vol. 24, No. 1, 2024, p. 220.

[9] I. Guyon, A. Elisseeff, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research, Vol. 3, 2003, pp. 1157-1182.

[10] Y. Li, T. Li, H. Liu, "Recent Advances in Feature Selection and Its Applications", Knowledge and Information Systems, Vol. 53, No. 3, 2017, pp. 551-577.

[11] A. Negi, V. Jaiswal, "A First Attempt to Develop a Diabetes Prediction Method Based on Different Global Datasets", Proceedings of the Fourth International Conference on Parallel, Distributed and Grid Computing, Waknaghat, India, 22-24 December 2016, pp. 237-241.

[12] N. P. Tigga, S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods", Procedia Computer Science, Vol. 167, 2020, pp. 706-716.

[13] UCI Machine Learning, "Pima Indians Diabetes Database", https://www.kaggle.com/uciml/pima-indians-diabetes-database (accessed: 2025)

[14] B. Strack et al. "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records", BioMed Research International, Vol. 2014, No. 1, 2014, pp. 1-11.

[15] M. Heydari, M. Teimouri, Z. Heshmati, S. M. Alavinia, "Comparison of Various Classification Algorithms in The Diagnosis of Type 2 Diabetes in Iran", International Journal of Diabetes in Developing Countries, Vol. 36, No. 2, 2016, pp. 167-173.

[16] M. M. F. Islam, R. Ferdousi, S. Rahman, H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques", Proceedings of the International Symposium on Computer Vision and Machine Intelligence in Medical Image Analysis, 2019, pp. 113-125.

[17] M. F. Dzulkalnine, R. Sallehuddin, "Missing Data Imputation with Fuzzy Feature Selection for Diabetes Dataset", SN Applied Sciences, Vol. 1, No. 4, 2019, p. 362.

[18] O. O. Oladimeji, A. Oladimeji, O. Oladimeji, "Classification Models for Likelihood Prediction of Diabetes at Early Stage Using Feature Selection", Applied Computing and Informatics, Vol. 20, No. 3/4, 2021, pp. 279-286.

[19] M. Abedini, A. Bijari, T. Banirostam, "Classification of Pima Indian Diabetes Dataset using Ensemble of Decision Tree, Logistic Regression and Neural Network", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 9, No. 7, 2020, pp. 1-4.

[20] A. Iyer, S. Jeyalatha, R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process, Vol. 5, No. 1, 2015, pp. 01-14.

[21] V. Chang, J. Bailey, Q. A. Xu, Z. Sun, "Pima Indians Diabetes Mellitus Classification Based on Machine Learning (ML) Algorithms", Neural Computing and Applications, Vol. 35, No. 22, 2023, pp. 16157-16173.

[22] K. Dashdondov, S. Lee, M.-U. Erdenebat, "Enhancing Diabetes Prediction and Prevention through Mahalanobis Distance and Machine Learning Integration", Applied Sciences, Vol. 14, No. 17, 2024, p. 7480.

[23] G. Pradhan et al. "Optimized Forest Framework with A Binary Multineighborhood Artificial Bee Colony for Enhanced Diabetes Mellitus Detection", International Journal of Computational Intelligence Systems, Vol. 17, No. 1, 2024, p. 194.

[24] G. Pradhan, G. Thapa, R. Pradhan, B. Khandelwal, S. Visalakshi, "A Study on Transcontinental Diabetes Datasets Using a Soft-Voting Ensemble Learning Approach", Proceedings of Advances in Communication, Devices and Networking, 2023, pp. 87-99.

[25] S. Habibi, M. Ahmadi, S. Alizadeh, "Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining", Global Journal of Health Science, Vol. 7, No. 5, 2015, pp. 304-310.

[26] S. Gündoğdu, "Efficient Prediction of Early-Stage Diabetes Using XGBoost Classifier with Random Forest Feature Selection Technique", Multimedia Tools and Applications, Vol. 82, No. 22, 2023, pp. 34163-34181.

[27] S. Buyrukoğlu, A. Akbaş, "Machine Learning based Early Prediction of Type 2 Diabetes: A New Hybrid Feature Selection Approach using Correlation Matrix with Heatmap and SFS", Balkan Journal of Electrical and Computer Engineering, Vol. 10, No. 2, 2022, pp. 110-117.

[28] N. Nipa, M. H. Riyad, S. Satu, Walliullah, K. C. Howlader, M. A. Moni, "Clinically adaptable machine learning model to identify early appreciable features of diabetes", Intelligent Medicine, Vol. 4, No. 1, 2024, pp. 22-32.

[29] B. P. Kumar, "Diabetes Prediction and Comparative Analysis Using Machine Learning Algorithms", International Research Journal of Modernization in Engineering Technology and Science, Vol. 04, No. 05, 2022, pp. 1-9.

[30] A. Yasar, "Data Classification of Early-Stage Diabetes Risk Prediction Datasets and Analysis of Algorithm Performance Using Feature Extraction Methods and Machine Learning Techniques", International Journal of Intelligent Systems and Applications in Engineering, Vol. 9, No. 4, 2021, pp. 273-281.

[31] J. Ma, "Machine Learning in Predicting Diabetes in the Early Stage", Proceedings of the 2nd International Conference on Machine Learning, Big Data and Business Intelligence, Taiyuan, China, 23-25 October 2020, pp. 167-172.

[32] A. Saboor, A. U. Rehman, T. M. Ali, S. Javaid, A. Nawaz, "An Applied Artificial Intelligence Technique For Early Prediction of Diabetes Disease", International Conference on Latest Trends in Electrical Engineering and Computing Technologies, 2022, pp. 1-6.

[33] S. N. Elsadek, L. S. Alshehri, R. A. Alqhatani, Z. A. Algarni, L. O. Elbadry, E. A. Alyahyan, "Early Prediction of Diabetes Disease Based on Data Mining Techniques", International Conference on Computational Intelligence in Data Science, Vol. 611, 2021, pp. 40-51.

[34] Nurjahan, M. A. T. Rony, Md. S. Satu, M. Whaiduzzaman, "Mining Significant Features of Diabetes through Employing Various Classification Methods", Proceedings of the International Conference on Information and Communication Technology for Sustainable Development, Dhaka, Bangladesh, 27-28 February 2021, pp. 240-244.

[35] S. M. M. Hasan, Md. F. Rabbi, A. I. Champa, Md. A. Zaman, "A Machine Learning-Based Model for Early Stage Detection of Diabetes", Proceedings of the International Conference on Computer and

Information Technology, Dhaka, Bangladesh, 19-21 December 2020, pp. 1-6.

[36] A. Mousa, W. Mustafa, R. B. Marqas, S. H. M. Mohammed, "A Comparative Study of Diabetes Detection Using The Pima Indian Diabetes Database", The Journal of University of Duhok, Vol. 26, No. 2, 2023, pp. 277-288.

[37] R. Rajni, A. Amandeep, "RB-Bayes Algorithm for The Prediction of Diabetic in Pima Indian Dataset", International Journal of Electrical and Computer Engineering, Vol. 9, No. 6, 2019, p. 4866.

[38] D. Sisodia, D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms", Procedia Computer Science, Vol. 132, 2018, pp. 1578-1585.