Article

Diagnosis of diabetes: A machine learning paradigm using optimized features

Rafid Mostafiz[1,2], **Khandaker Mohammad Mohi Uddin**[2], **Mohammad Shorif Uddin**[3], **Farhana Binte Hasan**[2], **Mohammad Motiur Rahman**[1]

[1]Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh

[2]Department of Computer Science and Engineering, Dhaka International University, Dhaka, Bangladesh

[3]Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh

E-mail:rafid.dka@gmail.com,jilanicsejnu@gmail.com,shorifuddin@gmail.com,farhana.cse.bd@gmail.com,mm73rahman@gmail.com

Received 24 April 2021; *Accepted 29 May 2021*; *Published 1 September 2021*

Abstract

Diabetes is considered one of the incurable diseases at present which is caused by hyperglycemia. Modern healthcare finds some attributes such as uncontrolled lifestyle, lack of balanced diets, genetic complexities, excess mental fatigue, obesities, and so on, which are responsible to precipitate the rapid mobility of diabetes diseases. This is not only a single disease but it also damages the nervous systems, heart, kidney, liver, eyes, and various organic metabolisms. Currently, the clinical industries have a huge amount of data for the diagnosis of diabetic patients. Machine learning algorithms can work appropriately to mitigate this tedious task in finding hidden patterns, discovering knowledge from the database, and predict outcomes. This research has proposed an efficient machine learning-based diagnosis methodology that outperforms the existing similar methodologies. The experiment selects the minimum Redundancy Maximum Relevance (mRMR) features from the working dataset and then recursive feature elimination (RFE) technique for optimization. The irregularity problem in the dataset is addressed by the synthetic minority oversampling technique (SMOTE). Machine learning classification is performed on the selected optimized features through Decision Tree (C4.5 DT), K-Nearest Neighbors (KNN), Naive Bayes (NBs), Support Vector Machine (SVM), Logistic Regression (LGR), and Random Forest (RF), where RF classifier produces best-suited results with minimum false detection rate. This experiment has used a 5-fold cross-validation approach to justify the reliability of the proposed model and finally obtain an accuracy of 98.10%.

Keywords diabetic; machine learning; minimum Redundancy Maximum Relevance; Recursive Feature Elimination; Random Forest Classifier.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Diabetes is one of the most talked-about issues in the world today. Diabetes is a silent killer. In today's world,

diabetes is not limited to one disease, but it can lead to kidney failure, heart disease, brain stroke, blindness, and even death. The human body releases a type of fluid called insulin, which works to produce energy. If for some reason insulin in the body loses its effectiveness which is called insulin resistance, then the amount of glucose in the blood increases, which leads to diabetes (NCD Risk Factor Collaboration, 2016). There are 3 types of diabetes: type 1 diabetes, type 2 diabetes, and gestational diabetes. Type 1 is more common in children, and teenagers, type 2 diabetes can occur at any age, and gestational diabetes can occur in women during pregnancy (Zimmet, 2016). According to the WHO's 2014 survey, there are about 422 million people worldwide with diabetes and by 2040 it will reach 642 million (WHO, 2021).

In today's world, medical diagnosis and clinical data analysis have been blessed with the computational reliability of artificial intelligence (AI). AI has expanded where human vision is limited. A variety of applications are being used in medical science now for data analysis and innovation through different machine learning algorithms. The use of machine learning algorithms has been seen in many recent healthcare studies, for example, MRI based tumor detection method (Mostafiz et al., 2021), prediction of heart diseases (Garg et al., 2021), polyp detection using endoscopic video frames (Mostafiz et al., 2020a, 2020b), liver lesion diagnosis (Mostafiz et al., 2020a), etc. Recently, in diabetes diagnosis machine learning and data mining methods are showing promising potentiality (Choubey et al., 2020; Haq et al., 2020).

Machine learning algorithms are efficient in reducing the misdetection rate and time complexity. With this view, here, for diabetes diagnosis, a dataset is formed using some basic input features, such as blood pressure, glucose level, skin thickness, BMI, insulin level, age, etc. Optimization of this dataset is investigated with the help of mRMR, DISR, CMIM, RFE, and SMOTE (Haq et al., 2020). The machine learning classifiers like DT, KNN, NB, SVM, LGR, RF are evaluated individually using the optimized feature vector. Experimental results are compared based on various measurement metrics to find the promising model. The distinct outcomes of this research are mentioned below:

- A computer-aided diagnosis method for diabetic-disease diagnosis is developed whose false diagnosis rate is low.
- An extensive analysis of feature optimization techniques is presented to select important feature values.
- A relative evaluation of different machine learning classifiers is explored to achieve the best-suited one.

For reading ease all the abbreviations used in this paper are given in Table 1. The remaining parts of the paper are divided into the following sections. Section 2 briefly describes the relative study of existing literature. Section 3 reflects the research methodology. Section 4 presents the experimental results and analyses. Finally, section 5 concludes the paper.

2 Related Works

A plethora of research works has been done to predict diabetes from several datasets using different algorithms and approaches. Polat and Gunes (Polat and Gunes, 2007) proposed a classification system where PCA is used in the first phase to decrease the dimensions and in the second phase, they used ANFIS (Adaptive Neuro-Fuzzy Inference System) to classify the obtained features from the first phase. LDA and ANFIS are combinedly used by Dogantekin et al. (Dogantekin et al., 2010) for diabetes classification. They divided their work into two stages: LDA is used in the first stage to isolate feature variables and ANFIS is used in the second stage for doing the classification.

Ali et al. (2015) did the performance analysis of several classification methods, such as SMO, KStar, Naive Bayes, AdaBoost, LMT, PART, J48, JRip, Random Tree, and OneRising multiple datasetsfrom the UCI repository. Sharma et al. (2015) also did similar work for the nearest neighbor (KStar), decision tree (M5P),

rule-based (M5 rule), neural network (multilayer perceptron), etc.

Guo et al. (2012) used the Bayesian network to predict type 2 diabetes using Pima Indian diabetes dataset. They found that their used architecture gave an accurate and efficient result. Similarly, Wu et al. (2018) used the logistic regression method and K-means clustering on Pima Indian diabetes dataset to predict type 2 diabetes mellitus. To classify diabetes, Kumar et al. (2017) used multilayer Perceptron, binary logistic regression, and KNN algorithms. The authors observed that KNN achieved better performance among other classification algorithms. In addition, other researchers (Maniruzzaman et al., 2020; Zou et al., 2018) also investigated different machine learning strategies for diabetes prediction.

Table 1 Abbreviations that are used in this paper.

AdaBoost	Adaptive Boosting
ANFIS	Adaptive Neuro-Fuzzy Inference System
AUC	Area Under Curve
CMIM	Conditional Mutual Information Maximization
CV	Cross-Validation
DISR	Double Input Symmetrical Relevance
DT	Decision Tree
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
KNN	K-neatest Neighbor
LDA	Linear Discriminate Analysis
LGR	Logistic Regression
ML	Machine Learning
MLP	Multilayer Perception
mRMR	Minimum redundancy and maximum relevance
NBs	Naïve Bayes
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
RF	Random Forest
RFE	Recursive Feature Elimination
SD	Standard Deviation
SMO	Sequential Minimal Optimization
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UCI	University of California, Irvine

From the above literature, it is found that the results of several machine learning algorithms are showing mixed and confusing results. To overcome this situation, we are motivated to work on diabetes diagnosis comprehensively.

3 Material and Methods

The working procedure of our methodology is demonstrated in Fig. 1. We have used the state-of-the-art datasets and these datasets are preprocessed for our exploratory analysis. There are 15 different features in the datasets, which are optimized to get 7 significant informative features. The proposed model investigated mRMR, DISR, and CMIM feature selection techniques. The RFE technique has been applied for the feature optimization from the selected feature vector. Then the machine learning (ML) classifier has been used to classify whether the diagnostic test is diabetic or not diabetic. The comparative analysis is performed to find the best model among several ML classifiers: Decision Tree (C4.5 DT), K-Nearest Neighbors (KNN), Naïve Bayes (NBs), Support Vector Machine (SVM), Logistic Regression (LGR), and Random Forest (RF).

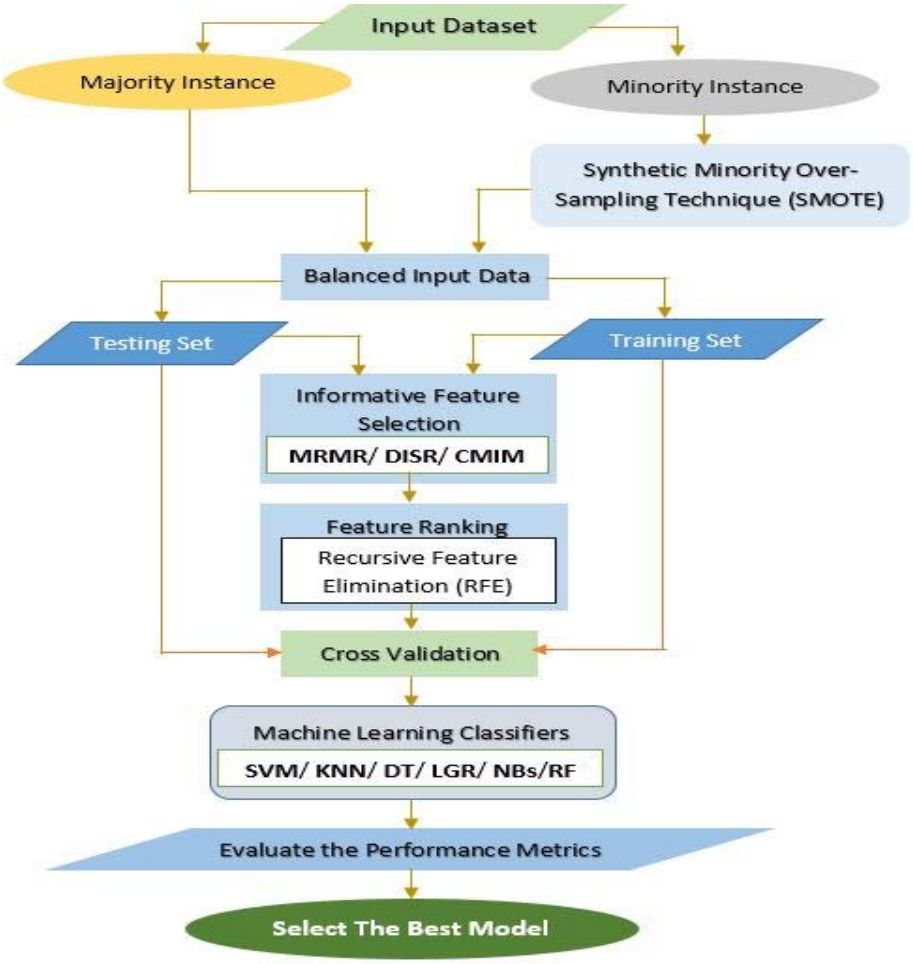

Fig. 1 Working procedure of the proposed experiment.

3.1 Experimental dataset

Total 6576 instances are collected from various data repositories (Choubey et al., 2020; Maniruzzaman et al., 2020; Zou et al., 2018). The dataset contains 2381 instances for diabetic positive data and 4195 instances for diabetic negative data. The total dataset is randomly split into training and testing sets at a ratio of 7:3. The distinct features of the data are listed in Table 2. The dataset shows that the positive instances are almost half of the negative instances. This skew in the data distribution is known as an imbalanced dataset which may

bias the diagnosis results. Thus, a data balancing scheme like, synthetic minority over-sampling technique (SMOTE) is applied on the feature set as a tricky solution. SMOTE creates new samples from the minority set by interpolating their neighborhood points. This technique will mitigate the overfitting problem than using random sampling whereas replace or increase the data randomly. Moreover, it shows very much effective outcomes in this high-dimensional dataset with no loss of informative features. SMOTE uses to create over-sampling synthetic examples by feature-space rather than over-sampling with replacement of dataspace (Das et. al., 2018). The synthetic example is chosen from the minority class sample in each over-sampling and joined along the line segment of all the *k* nearest neighbor of that minor class point. The neighbors are randomly taken based on the number of over-sampling is required denoted as *p*. The synthetic samples are considered as, the nearest difference between the feature vector samples and their neighbors. Then the difference is multiplied by a random value between 0 to 1 and added to the feature sample. This will help to select a point between two specific features. The working procedure of SMOTE in this experiment is shown by the flowchart of Fig. 2.

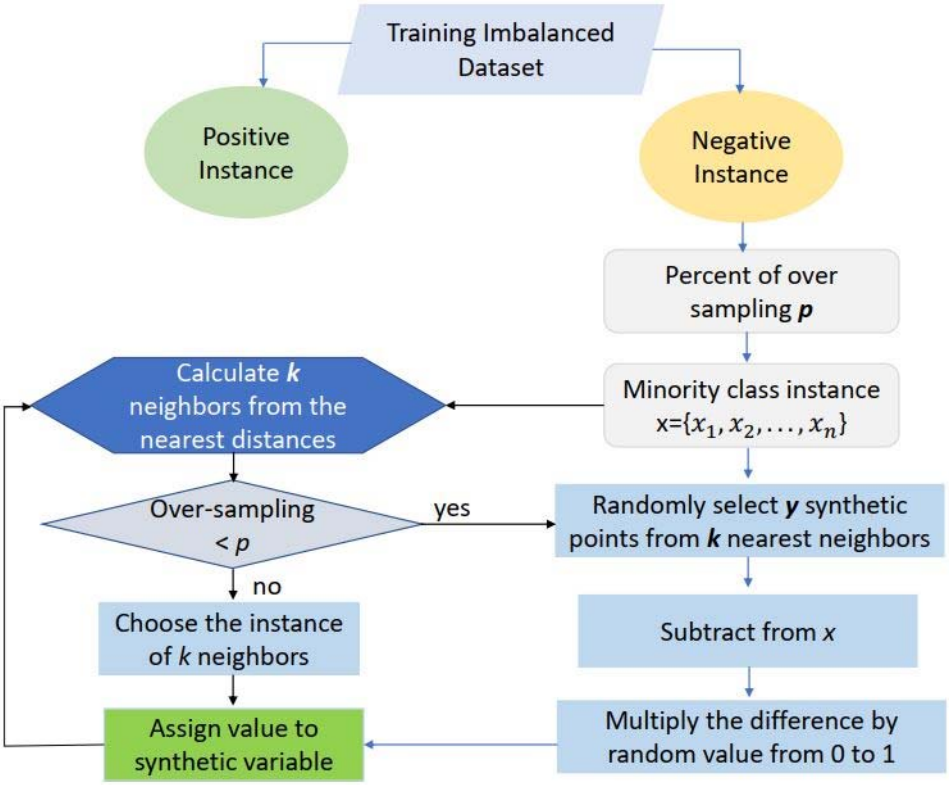

Fig. 2 Flowchart of SMOTE algorithm.

Table 2 Features description.

Features	Description
Age	Year ranges
Sex	Male/Female
Fasting blood sugar	Milligram per deciliter (mg/dL)
Diastolic blood pressure	Mille meter per mercury (mm

	Hg)
Systolic blood pressure	Mille meter per mercury (mm Hg)
Average blood sugar (HbA1C %)	Mille moles per mole (mmol/mol)
Waist circumference	Centimeter (Cm)
Hip circumference	Centimeter (Cm)
Total cholesterol	Milligram per deciliter(mg/dL)
Family history of diabetes	Yes/ No
Education	Yes/ No
Marital Status	Married/ Unmarried
Occupation	Office-work/ Field
Physical activity	Yes/ No
2 h post glucose loador Oral glucose tolerance test (OGTT)	Milligram per deciliter(mg/dL)

3.2 Machine learning algorithms

This segment presents a brief overview of the working procedure of ML classifiers used in this experiment. The classification methods utilize C4.5 DT, KNN, LGR, SVM, NBs, RF classifiers to find the best-suited classification model through a comparative analysis.

3.2.1 Naïve Bayes

Naïve Bayes (NBs) outperforms the complex classification method assuming the presence of a particular class in case of the absence of similar features. Using the Bayes' theorem NBs classifier is designed based on conditional probability (Siddique et al., 2013). NBs is using as a supervised ML technique in medical statistical data analysis as it is highly efficient on the diverse volume of the dataset as well as finding the meaningful conclusion. This classifier is mapped using equation 1, where the $P(H|D)$ is the probability input hypothesis for a given data, depending on the probabilities of $P(D|H)$, $P(H)$, $P(D)$.

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} \tag{1}$$

$P(D|H)$ observes the conditional probability distribution for each of the input instance $D= \{D_1, D_2, \ldots, D_n\}$ given for the response variable H. The marginal probability of the response variable is represented by $P(H)$. $P(D)$ is the marginal probability of input instance. This experiment computes the response variable both for H=*Positive* and H=*Negative* to obtain the probability of diabetic positive or negative, respectively. Equation 2 is derived to assign a function to predict the class label using NBs classifier.

$$H = arg\max_H P(H) \prod_{i=1}^{n} P(D_i|H) \tag{2}$$

3.2.2 Logistic Regression

Logistic Regression (LGR) works as a binary classifier in this experiment to predict the probability of diabetic positive or negative as the diagnosis result. The prediction is based on the value of the logit function for each input variable (Hosmer et al., 2013). According to the given value of the independent variables, it gives the

output as '0 or '1'. Basically, in case a probability score is smaller than 0.5, LGR will classify it as '0'; something else as '1'. For the data points of D= $\{D_1, D_2, \ldots, D_n\}$ it calculates the linear equation denoted by equation 4. The logit function using equation 5is used to squeeze the output of a linear equation between 0 and 1. Equation 3 calculates the maximum likelihood estimation for the regression coefficient W^T.

$$W^T = max \sum_{i=1}^{n} Y_i \times W_i D_i \tag{3}$$

$$z = Y_i \times W^T D_i \tag{4}$$

$$P(z) = \frac{1}{1+exp^{-z}} \tag{5}$$

The probability of input instance as diabetic positive is indicated while $P(z)$>0.5. Y_i indicates the data points.

3.2.3 Support Vector Machine

Support vector machine (SVM) is a well-known supervised learning technique in machine learning for data classification and regression. SVM performs classification by mapping each data item into a high-dimensional feature space. Classifying the data into two classes SVM creates a hyperplane. The best hyperplane is created by the SVM for the linear data by minimizing the marginal distance between two classes such as: {positive (+1) and negative (-1)} and minimizing the generalization errors (Mostafiz et al., 2020b). The input dataset denoted by D= $\{(D_1, y_1), (D_2, y_2) \ldots, (D_n, y_n)\}$ indicating the training instance with corresponding positive or negative classes $y_i$where $y_i \epsilon$ {+1 or -1}. The separating hyperplane is formed using equation 6 by calculating the distance maximization of $W^T D_i + b = -1$for y_i = diagnosis negative and $W^T D_i + b = +1$ for y_i= diagnosis positive. Then equation 7 indicates the accuracy of classification for those points satisfy it.

$$W^T D_i + b=0 \tag{6}$$

$$Y_i . W^T D_i + b \geq 1 \tag{7}$$

Here, the optimization function is denoted by $f(\dot{w}, \dot{D}_i)$ such that the value of $\frac{2}{\|w\|}$ should be maximum. It indicates the hyperplane with a larger margin of reciprocal magnitude as equation 8.

$$(\dot{w}, \dot{D}_i) = \min \frac{\|w\|}{2} + C_i . \sum_{i=1}^{k} \lambda_i . \tag{8}$$

For better performance, the optimum choice of the kernel parameter λ and regularization parameter C is essential.

3.2.4 K-Nearest Neighbors

For doing classification and solving regression problems k-nearest neighbor (KNN) is the simplest machine learning approach (Hossain et al., 2019). In KNN, sample points find out the smallest distance of all the points of the training dataset. Whenever any test instant (x, y) is put into the training data plane D= $\{(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)\}$, Euclidean distance function is used to find the distanced_i from all the data points of D according to equation 9. A first kth minimum distance of d_i is pushed in the vector r_i denoted by equation 10.

When KNN works as a classifier the output *S* will be a class member that belongs to the majority counts. The class level '+1' indicates the positive diagnosis and '-1' indicates the negative diagnosis result according to equation 11.

$$d_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}; \qquad i = 1 \ldots n \tag{9}$$

$$r_i = min_{i=1}^{k} d_i \tag{10}$$

$$S = \max\left(\text{count}_{,r\epsilon[+1]}\, \text{count}_{r\epsilon[-1]}\right) \tag{11}$$

3.2.5 Decision Tree (C4.5)

A decision tree (DT) actually builds a tree where nodes represent input features. The output is then predicted based on the height information gain for that particular feature. The subtree is formed using the features that are not utilized in the over steps. The internal nodes represent input variables, the branch represents outcomes, and leaf nodes represent classes (Nookala et al., 2013). DT consists of multiple levels of nodes - the top-most node known as root node and leaves are the least level nodes.

The entropy of each class is calculated by equation 12 and the entropy depending on two features is then obtained by equation 13. These are utilized to find the feature-level gain *G*from equation 14.

$$H(S) = \sum_{i=1}^{c} -P_i \log_2 P_i \tag{12}$$

$$H(S, D) = \sum_{c\epsilon D} P(c).E(c) \tag{13}$$

$$G(S, D) = H(S) - H(S, D) \tag{14}$$

Here, the entropy of the leaf node is 0 and the C4.5 DT follows this condition to recursively visit the non-leaf node until classifying the input data. The data features are split consecutively and evaluate the entropy of each branch to sum up the total proportional entropy*I*by using equation 15. The height gain ratio of the feature level is obtained by *G*/*I*.

$$I(S, D) = -\sum_{j=1}^{v} \frac{D_i}{D} \log_2 \frac{D_i}{D} \tag{15}$$

3.2.6 Random Forest

Random forest (RF) consists of many DTs and it is a state-of-art classification algorithm (Breiman, 2001). For classifying a new data sample, the input feature vector is passed to all the DTs of that RF. Each DT does the classification based on the input feature. The final classification output is considered by the outcome of the DT who gets the most 'votes' (Mostafiz et al., 2020d). The graphical illustration of the RF classifier is presented in Fig. 3, where '0' indicates a negative diabetic case and '1' indicates a positive diabetic case.

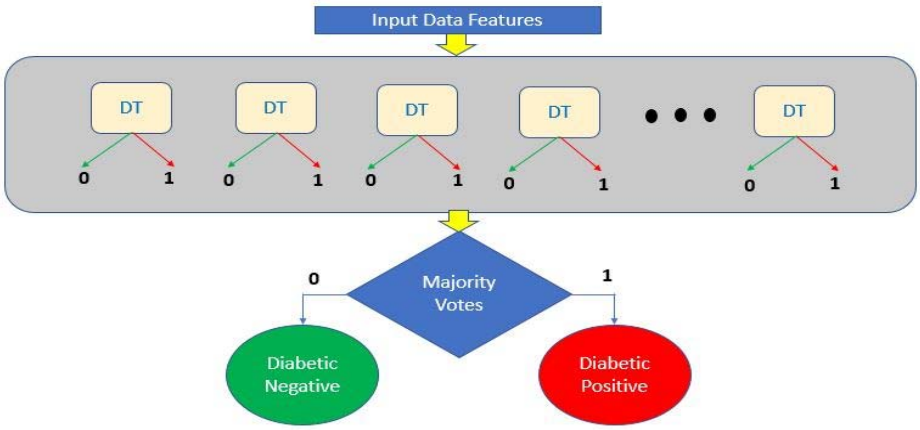

Fig. 3 Schematic diagram of random forest (RF) classifier.

RF (Shah et al., 2019) follows the following steps:

- Dataset is separated into two parts: training and testing. Using the bootstrapping method, the training set creates a new dataset.
- Based on the result of step 1 DT is constructed.
- After repeating step 1 and step 2 many trees are produced which make a forest.
- For the input variables, every tree in the forest gives a vote.
- The votes for each class are computed. For the input variables, the classifier sets the class which gets the highest votes.

3.3 Feature optimization

In machine learning-based classification, pragmatic feature selection is very crucial to establish more accurate results with minimum error. The exploratory analysis has found that irrelevant and redundant features seriously degrade the performance of theclassifier. This research has focused on mutual informative feature selection techniques such as Conditional Mutual Information Maximization (CMIM) (Fleuret and Francois, 2004), Double Input Symmetrical Relevance (DISR) (Meyer et al., 2008), and minimum Redundancy Maximum Relevance (mRMR) (Peng et al., 2005) to select the best subset of feature variables.

3.3.1 Conditional Mutual Information Maximization (CMIM)

CMIM calculates the conditional dependencies between two feature variables for a given third variable. It selects the features whose conditional relevance is minimal among the prior selected features. For a given input variable, it maximizes the distance of the mutual feature variables. If the selected variable is highly complementary with already selected variables, then it is characterized as a high conditional mutual information with that variable (Liang et al., 2019). Equation 16 indicates the relevance feature selection with no redundancy.

$$X_{CMIM} = arg\ \max_{x_i \epsilon D} \{\min_{X \in S_X} I(x_i; y|X)\} \tag{16}$$

The input feature variable x_i belongs to the instance *D* if the mutual dependency $I(x_i; y|X)$ is minimal among the prior selected feature set S_X. The output *y* requires the maximal minimum conditional dependencies based on the mutual information between $x_i \epsilon D$ and $X \in S_X$.

3.3.2 Double Input Symmetrical Relevance (DISR)

DISR considers the pair of feature variables whose combination returns more information than the sum of their individual feature on the target variable. This will find a feature set *S* of *n* variables such that it will maximize the mutual information for target variable *T*. The most promising subset computes the highest average sum of mutual information among all the possible combinations of two variables (double input). This is called the symmetrical relevance of all the combinations. For a given two random variables*x* and *y*, the symmetrical relevance is measured by equation 17.

$$R(x,y) = \frac{I(x,y)}{H(x,y)} \tag{17}$$

The resulting criterion of selecting relevant variables along with avoiding redundancy can be computed by equation 18.

$$X_{DISR} = arg\max_{x_i \epsilon D} \left\{ \sum_{x_j \epsilon S} R(x_{i,j}, T) \right\} \tag{18}$$

3.3.3 Minimum Redundancy Maximum Relevance (mRMR)

Minimum redundancy maximum relevance (mRMR) (Wang et al., 2018) guarantees the pairwise features have minimized correlations and have the maximum Euclidean distances. It is a feature optimization algorithm. For the selected features it optimizes the mutual dependencies. Equation 19 is used to calculate the mutual dependencies of two variables *x* and *y*.

$$I(x,y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{19}$$

$$maxD(S,c) = \frac{1}{|S|} \sum_{x_i \in c} I(x_i; c) \tag{20}$$

$$minR(S) = \frac{1}{|S|^2} \sum_{x_i x_j \in c} I(x_i; x_j) \tag{21}$$

D(*S*, *c*) denotes maximal relevance can be calculated using equation 20. Here, x_i is the mean of all mutual dependencies and S represents the feature set, $R(S)$ is the minimal redundancies and calculated by equation 21.

$$max\varphi(D,r);\ \ \varphi = D - R \tag{22}$$

Equation 22 is used to obtain a good subset of features by optimizing the relevance and redundancy.

3.4 RFE feature elimination (RFE)

The machine learning classification may not produce the desired outcomes with selected features (Chen et al., 2018). For this, RFE warps the features and finds the optimized feature set. It performs elimination of the highly correlated features that contain the same information. The obtained feature set avoids redundant features based on the predictor scores. Using RFE the least important predictors are removed recursively by re-computing the feature ranking. In this experiment, the dataset contains a number of rows and columns where columns denote features and rows denote samples.The feature variablesare ranked by passing through the ML classifier and achieved the optimized target set. A cross-validation (CV) approach is performed while ranking the variablesthrough RFE. In the exploratory analysis, RFE identifies the best features by eliminating the less important and redundant featureswith the steps of cross-validation. In this research, the optimized features are ranked using RFE, based on the accuracy and correlation values.Its working steps are as follows:

- The input feature set S from the feature selection scheme is warped by RFE.

- The lower rank features are rescued based on the correlation and accuracy metrics.
- The highly correlated feature set R is excluded from the prior selected set S, and get the optimized feature set SR=S-R.

The highly correlated features are bound not to interact with each other and evaluate their performance in different iterations. If their performance differs from each other and the difference exceeds a certain boundary, then the best-suited set is selected.

4 Results and Analysis

Our total 6576 data of the dataset are divided randomly into 4603 instances (about 70%) for training while 1973 instances (about 30%) for testing to evaluate the performance. The performance is measured using the terms of Accuracy, Precision, Recall, F-score, TPR, and FPR obtained from the confusion matrix. These metrics are defined through equations 23, 24, 25, 26 ,27, 28.

The area under the curve (AUC) is a statistical parameter that is beneficial to compare different machine learning algorithms based on TPR and the FPR. The measure of AUC confirms the discrimination capacity between diabetic positive and negative classes for the corresponding algorithm.

This experiment selects 7 features out of 15 using the feature selection techniques (mRMR, DISR, CMIM are done separately using RFE optimization). The feature selection has been conducted by focusing on the mutual correlation and then the RFE feature ranking is performed based on the upper bound and lower bound of the correlated values. Fig. 4 depicts the heat map visualizing the mutual correlation of the selected feature set obtained by mRMR feature selection and RFE optimization. This indicates that the mutual correlation among features is low. Table 3 presents the overview of the optimized feature set.

Fig. 4 Heat map of selected feature set.

Table 3 Experiment data with optimized feature value.

Features	Mean (mRMR/DISR/CMIM)	SD (mRMR/DISR/CMIM)	Min/Max
Glucose	120.9/115.6/122.5	31.9/28.5/35.8	0/199
Blood Pressure	69.1/67.5/68.9	19.4/15.2/22.3	0/122
Skin Thickness	20.5/20.9/19.2	15.9/16.2/18.6	0/99
Insulin	79.8/75.1/69.6	115.2/117.7/112.3	0/846
BMI	32.0/31.9/29.2	7.9/8.1/5.6	0/67
Diabetes Pedigree Function	0.47/0.51/0.46	0.33/0.41/0.39	0.078/2.42
Age	33.2/37.0/31.4	11.8/13.9/8.8	21.0/81.0

The performance is analyzed and observed in three phases. Firstly, the ML classifiers are evaluated on the working dataset without the feature optimization technique. Secondly, the optimization techniques (mRMR/ DISR/ CMIM) are applied with RFE on the working dataset and obtained the optimized feature values.The performances ofthe investigated ML classifiers (C4.5 DT, KNN, LGR, SVM, NBs, RF)are recorded separately. Finally, the best-suited model is proposed by evaluating the comparative analysis performance metrics.

$$Accuracy\ (\%) = \frac{TP+TN}{FP+TP+FN+TN} \times 100 \tag{23}$$

$$Recall\ (\%) = \frac{TP}{TP+FN} \times 100 \tag{24}$$

$$Precision\ (\%) = \frac{TP}{TP+FP} \times 100 \tag{25}$$

$$F-score\ (\%) = \frac{2\times Recall\times Precision}{Recall+Precision} \times 100 \tag{26}$$

$$TPR\ (\%) = \frac{TP}{TP+FN} \times 100 \tag{27}$$

$$FPR\ (\%) = \frac{FP}{FP+TN} \times 100 \tag{28}$$

Where TP, TN, FP, FN and TPR, and FPR are true positive, true negative, false positive, false negative, true-positive rate, and false-positive rate, respectively.The ROC (Receiver Operating Characteristics) curve is a plotting of TPR vs. FPR. The AUC (Area Under the ROC Curve) interprets the total prediction performance of a model. If a model's predictions are 100% wrong, then AUC is zero, if its predictions are 100% right that AUC is 1.0 in a normalized situation.

In the confusion matrix, the actual class indicates the input test instance and the predicted class indicates the outcome of the prediction. The 'Y_P' and 'N_N' in test results indicate diabetic positive and diabetic negative cases, respectively. The individual classification performance of ML classifiersis presented in Table

4 (confusion matrix) and Table 5 (score of the performance metrics) before feature optimization. Similarly, Table 6 (confusion matrix) and Table 7 (score of the performance metrics) using optimized features through CMIM and RFE, Table 8 (confusion matrix) and Table 9 (score of the performance metrics) using optimized features through DISR and RFE, Table 10 (confusion matrix) and Table 11 (score of the performance metrics) using optimized features through mRMR and RFE.

Table 4 Confusion matrix for individual classifier before feature optimization.

Classifiers	Actual Class (Y_P = 827; N_N = 1146)	Predict Class	
		Y_P	N_N
KNN	Y_P	(TP) 618	(FP) 209
	N_N	(FN) 157	(TN) 989
SVM	Y_P	(TP)688	(FP) 139
	N_N	(FN) 170	(TN) 976
C4.5 DT	Y_P	(TP) 701	(FP) 126
	N_N	(FN) 149	(TN) 997
LGR	Y_P	(TP) 723	(FP) 104
	N_N	(FN) 148	(TN) 998
NBs	Y_P	(TP) 768	(FP) 59
	N_N	(FN) 179	(TN) 967
RF	Y_P	(TP) 745	(FP) 82
	N_N	(FN) 129	(TN) 1017

Table 5 Evaluation of classification methods before feature optimization.

Classifiers	Accuracy	Recall	Precision	F-score
KNN	0.8146	0.7974	0.7479	0.7719
SVM	0.8434	0.8019	0.8319	0.8166
C4.5 DT	0.8602	0.8247	0.8476	0.8359
LGR	0.8719	0.8301	0.8766	0.8527
NBs	0.8793	0.8111	0.9287	0.8659
RF	0.8931	0.8524	0.9008	0.8759

Table 6 Confusion matrix for individual classifier using CMIM_RFE feature optimization.

Classifiers	Actual Class (Y_P = 827; N_N = 1146)	Predict Class	
		Y_P	N_N
SVM	Y_P	(TP) 738	(FP) 89
	N_N	(FN) 64	(TN) 1082
KNN	Y_P	(TP) 762	(FP) 65
	N_N	(FN) 98	(TN) 1048
NBs	Y_P	(TP) 766	(FP) 61
	N_N	(FN) 77	(TN) 1071
LGR	Y_P	(TP) 782	(FP) 45
	N_N	(FN) 90	(TN) 1056
C4.5 DT	Y_P	(TP) 771	(FP) 56
	N_N	(FN) 82	(TN) 1064
RF	Y_P	(TP) 769	(FP) 58
	N_N	(FN)66	(TN) 1080

Table 7 Evaluation of classification methods using CMIM_RFE feature optimization.

Classifiers	Accuracy	Recall	Precision	F-score
SVM	0.9224	0.9202	0.8924	0.9061
KNN	0.9174	0.9214	0.8861	0.9034
NBs	0.9311	0.9087	0.9262	0.9174
LGR	0.9316	0.8968	0.9456	0.9206
C4.5 DT	0.9301	0.9039	0.9324	0.9179
RF	0.9372	0.9211	0.9298	0.9254

Table 8 Confusion matrix for individual classifier using DISR_RFE feature optimization.

Classifiers	Actual Class (Y_P = 827; N_N = 1146)	Predict Class	
		Y_P	N_N
SVM	Y_P	(TP) 748	(FP) 89
	N_N	(FN) 61	(TN) 1085
KNN	Y_P	(TP) 779	(FP) 48
	N_N	(FN) 89	(TN) 1056

NBs	Y_P	(TP) 783	(FP) 44
	N_N	(FN) 90	(TN) 1056
LGR	Y_P	(TP) 769	(FP) 58
	N_N	(FN) 65	(TN) 1081
C4.5 DT	Y_P	(TP) 782	(FP) 45
	N_N	(FN) 47	(TN) 1099
RF	Y_P	(TP) 795	(FP) 32
	N_N	(FN)29	(TN) 1117

Table 9 Evaluation of classification methods using DISR_RFE feature optimization.

Classifiers	Accuracy	Recall	Precision	F-score
SVM	0.9241	0.9246	0.8937	0.9089
KNN	0.9301	0.8975	0.9421	0.9193
NBs	0.9321	0.8969	0.9468	0.9212
LGR	0.9378	0.9221	0.9297	0.9259
C4.5 DT	0.9534	0.9433	0.9456	0.9445
RF	0.9691	0.9648	0.9613	0.9630

Table 10 Confusion matrix for individual classifier using mRMR_RFE feature optimization.

Classifiers	Actual Class (Y_P = 827; N_N = 1146)	Predict Class	
		Y_P	N_N
KNN	Y_P	(TP) 759	(FP) 68
	N_N	(FN) 37	(TN) 1109
SVM	Y_P	(TP) 788	(FP)39
	N_N	(FN) 66	(TN) 1080
NBs	Y_P	(TP) 757	(FP) 60
	N_N	(FN) 25	(TN) 1121
LGR	Y_P	(TP) 783	(FP) 44
	N_N	(FN)48	(TN) 1098
C4.5 DT	Y_P	(TP) 808	(FP) 19
	N_N	(FN) 39	(TN) 1107
RF	Y_P	(TP) 809	(FP) 18
	N_N	(FN) 21	(TN) 1125

Table 11 Evaluation of classification methods using mRMR_RFE feature optimization.

Classifiers	Accuracy	Recall	Precision	F-score
KNN	0.9467	0.9535	0.9178	0.9353
SVM	0.9468	0.9293	0.9525	0.9408
NBs	0.9518	0.9680	0.9266	0.9468
LGR	0.9534	0.9422	0.9468	0.9445
C4.5 DT	0.9706	0.9541	0.9771	0.9655
RF	0.9810	0.9747	0.9782	0.9764

Tables 4 and Table 5 have confirmed that the RF gives the highest performance among the six machine learning classifiers without any feature selection and optimization. Tables 6 to Table 11 have confirmed that the RF gives the highest performance among the six machine learning classifiers with feature selection and optimization operations. These Tables also confirm that among three investigated feature selection strategies mRMR gives the best performance.We have found the highest accuracy of the RF method 98.10% through feature selection and optimization operations and 89.31% accuracy without feature selection and optimization. Therefore, it is confirmed that the feature selection along with optimization gives a large performance enhancement.The experiments have used a 5-fold cross-validation approach to justify the reliability of the proposed model. We have used a personal computer (PC) with an Intel Core i5 processor of 2.80 GHz, 12GB RAM, GEFORCE RTX 2070 GPU using 64-bit Windows 10 for conducting our work.

Several recent works of literature have focused on diabetic prediction using machine learning modalities. These related articles found some drawbacks like the inconsistent feature dimensions and data imbalance. Although there is not adequate literature to address these limitations, we have tried to compare our research outcomes with some related works based on a similar dataset. Table 12 shows such comparative performance. It also confirms the superiority of our method with a big margin.

Table 12 Comparison of our work with the most related works.

Paper	**Number of Instances**	**Feature Optimization**	**Classifier**	**Accuracy**	**AUC**
Sisodia et al., 2018	768	-	Naïve Bayes	76.30	0.819
Zou et al., 2018	178131	PCA/ mRMR	Random Forest	80.84	-
Ahuja et al., 2019	768	LDA	SVM	91.6	-
Mujumdar et al., 2019	800	K-means	Logistic Regression	96	-
Semerdjian et al., 2017	5515	Random Forest	Gradient Boosting	-	0.84
Mohapatra et al., 2019	768	-	MLP	77.50	-

Yu et al., 2010	6314	-	SVM	83.50	0.834
Pei et al., 2018	10436	Cross Selection	J48 DT	94.2	0.948
Choubey et al., 2020	1058	PCA	C4.5 DT	95.58	0.981
Maniruzzaman et al., 2020	6561	Logistic Regrassion	Random Forest	94.25	0.95
Proposed	6576	mRMR + RFE	Ensemble Random Forest	98.10	0.998

5 Conclusion

The main aim of this research is to develop an efficient model for diabetic disease diagnosis using clinical data with optimized features. Data preprocessing was performed using SMOTE to overcome the data imbalance problem. This research has found that accurate feature selection plays a dominant role in automatic diagnosis. mRMR feature selection along with RFE optimization gives the best performance in selecting seven important features for the diabetes diagnosis. Extensive experimentation is performed using six machine learning models. RF classifier is the best model using mRMR features selection and RFE optimization technique with an accuracy of 98.10% and an AUC of 0.998. This is really an excellent outcome outperforming the existing methods. In the future, the work can be extended to analyze the statistical data of other clinical modalities in a machine learning fashion.Besides, successful treatment and recovery prediction rate analysis may also be an option for further study.

Acknowledgment

We are thankful to the venues of all the open source dataset.

References

Ahuja R, Vivek V, Chandna M, Virmani S, Banga A. 2019. Comparative study of various machine learning algorithms for prediction of insomnia. In: Advanced Classification Techniques for Healthcare Analysis. 234-257, IGI Global

Ali FM, Fgee EBE, Zubi ZS. 2015. Predicting performance of classification algorithms. International Journal of Computer Aided Engineering and Technology, 6(2): 19-28

Breiman L. 2001. Random forests. Machine Learning, 45(1): 5-32

Choubey DK, Kumar P, Tripathi S, Kumar S. 2020. Performance evaluation of classification methods with PCA and PSO for diabetes. Network Modeling Analysis in Health Informatics and Bioinformatics, 9(1): 1-30

Chen Q, Meng Z, Liu X, Jin Q, Su R. 2018. Decision variants for the automatic determination of optimal feature subset in RF-RFE. Genes, 9(6): 301

Das S, Datta S, Chaudhuri BB. 2018. Handling data irregularities in classification: Foundations, trends, and future challenges. Pattern Recognition, 81: 674-693

Dogantekin E, Dogantekin A, Avci D, Avci L. 2010. An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS. Digital Signal Processing, 20(4): 1248-1255

Fleuret F. 2004. Fast binary feature selection with conditional mutual information. Journal of Machine

Learning Research, 5(9)

Garg A, Sharma B, Khan R. 2021. Heart disease prediction using machine learning techniques. In: IOP Conference Series: Materials Science and Engineering. 1022(1): 012046, IOP Publishing, UK

Guo Y, Bai G, Hu Y. 2012. Using Bayes network for prediction of type-2 diabetes. In: 2012 International Conference for Internet Technology and Secured Transactions. 471-472, IEEE

Haq AU, Li JP, Khan J, Memon MH, Nazir S, Ahmad S, Khan GA, Ali A. 2020. Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data. Sensors, 20(9): 2649

Hosmer Jr DW, Lemeshow S, Sturdivant RX. 2013. Applied Logistic Regression (Vol. 398). John Wiley and Sons, USA

Hossain ME, Khan A, Moni MA, Uddin S. 2019. Use of electronic health data for disease prediction: A comprehensive literature review. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18(2): 745-758

Liang J, Hou L, Luan, Z, Huang W. 2019. Feature selection with conditional mutual information considering feature interaction. Symmetry, 11(7): 858

Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. 2020. Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8(1):1-14

Meyer PE, Schretter C, Bontempi G. 2008. Information-theoretic feature selection in microarray data using variable complementarity. IEEE Journal of Selected Topics in Signal Processing, 2(3): 261-274

Mohapatra SK, Swain JK, Mohanty MN. 2019. Detection of diabetes using multilayer perceptron. In: International Conference on Intelligent Computing and Applications. 109-116, Springer, Singapore

Mostafiz R, Hasan M, Hossain I, Rahman MM. 2020a. An intelligent system for gastrointestinal polyp detection in endoscopic video using fusion of bidimensional empirical mode decomposition and convolutional neural network features. International Journal of Imaging Systems and Technology, 30(1): 224-233

Mostafiz R, Rahman MM, Uddin MS. 2020b. Gastrointestinal polyp classification through empirical mode decomposition and neural features. SN Applied Sciences, 2: 1-10

Mostafiz R, Rahman MM, Islam AKM, Belkasim S. 2020c. Focal liver lesion detection in ultrasound image using deep feature fusions and super resolution. Machine Learning and Knowledge Extraction, 2(3): 172-191

Mostafiz R, Uddin MS, Reza MM, Rahman MM., 2020d. Covid-19 detection in chest X-ray through random forest classifier using a hybridization of deep CNN and DWT optimized features. Journal of King Saud University-Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.12.010

Mostafiz R, Uddin, MS, Alam NA, Hasan MM, Rahman MM. 2021. MRI-based brain tumor detection using the fusion of histogram oriented gradients and neural features. Evolutionary Intelligence, 14(2): 1075-1087

Mujumdar A, Vaidehi V. 2019. Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165: 292-299

NCD Risk Factor Collaboration, 2016. Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19· 2 million participants. The Lancet, 387(10026):1377-1396

Nookala GKM, Pottumuthu BK, Orsu N, Mudunuri SB. 2013. Performance analysis and evaluation of different data mining algorithms used for cancer classification. International Journal of Advanced Research in Artificial Intelligence, 2(5)

Pei D, Zhang C, Quan Y, Guo Q. 2019. Identification of potential type II diabetes in a Chinese population with a sensitive decision tree approach. Journal of Diabetes Research, 2019: 4248218

Peng H, Long F, Ding C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8): 1226-1238

Polat K, Güneş S. 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing, 17(4): 702-710

Selvakumar S, Kannan KS, GothaiNachiyar S. 2017. Prediction of diabetes diagnosis using classification based data mining techniques. International Journal of Statistics and Systems, 12(2): 183-188

Semerdjian J, Frank S. 2017. An ensemble classifier for predicting the onset of type II diabetes. arXiv preprint, arXiv:1708.07480

Shah S, Luo X, Kanakasabai S, Tuason R, Klopper G. 2019. Neural networks for mining the associations between diseases and symptoms in clinical notes. Health information science and systems, 7(1): 1-9

Sharma R, Kumar S, Maheshwari R. 2015. Comparative analysis of classification techniques in data mining using different datasets. International Journal of Computer Science and Mobile Computing, 4(12): 125-134

Siddique AQ, Hossain MS. 2013. Predicting heart-disease from medical data by applying naive bayes and Apriori algorithm. International Journal Of Scientific and Engineering Research, 4(10): 224-231

Sisodia D, Sisodia DS. 2018. Prediction of diabetes using classification algorithms. Procedia Computer Science, 132: 1578-1585

Wang SP, Zhang Q, Lu J, Cai YD. 2018. Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm. Current Bioinformatics, 13(1): 3-13

World Health Organization. 2021. Diabetes. https://www.who.int/westernpacific/health-topics/diabetes Accessed Apr 21, 2021

Wu H, Yang S, Huang Z, He J, Wang X. 2018. Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10: 100-107

Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. 2010. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making, 10(1): 1-7

Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. 2018. Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9: 515

Zimmet P, Alberti KG, Magliano DJ, Bennett, PH. 2016. Diabetes mellitus statistics on prevalence and mortality: facts and fallacies. Nature Reviews Endocrinology, 12(10): 616-622