Bulletin of Electrical Engineering and Informatics
Vol. 12, No. 3, June 2023, pp. 1728~1737
ISSN: 2302-9285, DOI: 10.11591/eei.v12i3.4412

Comparative analysis of predictive machine learning algorithms for diabetes mellitus

Kirti Kangra, Jaswinder Singh
Department of Computer Science and Engineering, Faculty of Engineering and Technology, Guru Jambheshwar University of Science and Technology, Hisar, India

Article Info

Article history:

Received Jul 15, 2022
Revised Sep 30, 2022
Accepted Nov 27, 2022

Keywords:

Diabetes mellitus
Logistic regression
Machine learning
Support vector machine
WEKA

ABSTRACT

Diabetes mellitus (DM) is a serious worldwide health issue, and its prevalence is rapidly growing. It is a spectrum of metabolic illnesses defined by perpetually increased blood glucose levels. Undiagnosed diabetes can lead to a variety of problems, including retinopathy, nephropathy, neuropathy, and other vascular abnormalities. In this context, machine learning (ML) technologies may be particularly useful for early disease identification, diagnosis, and therapy monitoring. The core idea of this study is to identify the strong ML algorithm to predict it. For this several ML algorithms were chosen i.e., support vector machine (SVM), Naïve Bayes (NB), K nearest neighbor (KNN), random forest (RF), logistic regression (LR), and decision tree (DT), according to studied work. Two, Pima Indian diabetic (PID) and Germany diabetes datasets were used and the experiment was performed using Waikato environment for knowledge analysis (WEKA) 3.8.6 tool. This article discussed about performance matrices and error rates of classifiers for both datasets. The results showed that for PID database (PIDD), SVM works better with an accuracy of 74% whereas for Germany KNN and RF work better with 98.7% accuracy. This study can aid healthcare facilities and researchers in comprehending the value and application of ML algorithms in predicting diabetes at an early stage.

Corresponding Author:

Kirti Kangra
Department of Computer Science and Engineering, Faculty of Engineering and Technology
Guru Jambheshwar University of Science and Technology
Hisar, Haryana, India
Email: Kirtikangra98@gmail.com

1. INTRODUCTION

Today, the world is facing a lot of chronic diseases such as heart disease, cancer, diabetes, and tuberculosis. The early detection of these illnesses is crucial. The patient must endure these diseases for a very long time. Numerous studies are being done to control these diseases. But these diseases are becoming more prevalent day by day. More research is required to control these diseases. This paper will examine diabetes mellitus (DM), one of the chronic diseases.

DM usually known as diabetes, is a metabolic disorder marked by high blood sugar levels. In this insulin moves sugar from the bloodstream into cells and is accumulated or utilized to form energy. In the condition of diabetes patient’s body is not able to generate sufficient insulin or stop producing insulin. Chronic DM poses a number of health concerns and issues for humans. Type-1, type-2, pre-diabetes, and gestational diabetes are the most prevalent varieties of DM. Type-1 diabetes is a chronic disorder in which the patient's immune system assaults and abolishes the beta cells in the pancreas that secrete insulin. In type-2 diabetes, the body's insulin secretion diminishes, resulting in high blood sugar levels. According to recent studies, early identification can prevent 80% of type-2 diabetes. Pre-diabetes is a condition in which blood sugar levels are significant but not excessive enough to be diagnosed as type-2 diabetes. Pregnant women with excessive blood sugar are diagnosed with gestational diabetes [1].

Diabetes should be addressed as soon as possible because it can cause a slew of consequences. Diabetes is a severe disease; therefore, an automated approach to diabetes identification is critical. In the field of medicine, machine learning (ML) is popular because of the use of its algorithms that will improve the accuracy rate of disease detection and diagnosis [2]. Initial decisions at medical centers are based on the doctors' opinions and competence instead of the amount of hidden data in EHRs. It will lead to unintentional prejudices, errors, and unnecessary costs for patient treatment. To make diagnosis less costly and more accurate there is a need for ML. Therefore, using ML to predict diabetes can help doctors diagnose patients more efficiently and precisely.

a. Statistics of diabetes in India

India has an estimated 77 million diabetics, making it the world's second-largest diabetic population behind China. India has one in every six persons (17%) with diabetes in the world and contains 17.5% population (India) of the world as per the calculation of October 2018. According to TFPR editorial [3], the number will rise to 134 million by 2045. Among the total population of India, 72.96 million cases of diabetes are from the adult population (above 20 years). Among these patients, 10.9-14.2% are from urban areas and 3.0-7.8% are from rural areas [4].

b. Objective

By glancing at the statistics, it is clear that diabetes is a serious issue for the world and much more research is needed in this area. The key aim of this research is to identify a strong prediction algorithm from existing ML algorithms by analyzing research questions (RQs), that can assist hospitals and healthcare organizations. This paper can serve as a reference point for researchers interested in diabetes diagnosis. The following are the RQs: i) RQ1: what are the most widely used algorithms for diabetes prediction based on considerable research? and ii) RQ2: which of these algorithms performs better when it comes to performance analysis? To answer the discussed RQs, we must go through the related work. The related work will be covered in the next section.

2. RELATED WORK

Numerous work has already been done and is still going on. But there is further scope for improvement in the prediction of DM. For the answer to RQ1, a number of articles have been taken from different sources as IEEE Xplore, Science Direct, Google Scholar, and Research Gate. Ranging from 2019-2022. This section of the paper will discuss different techniques that are divided into traditional ML techniques and hybrid ML techniques.

2.1. Traditional ML technique

According to Mushtaq *et al.* [5] proposed a system to predict diabetes using ML algorithms at two stages. In the first stage, the dataset was balanced using synthetic minority oversampling technique (SMOTE), Tomek, and IQR. For classification at the first stage support vector machine (SVM), Naïve Bayes (NB), K nearest neighbor (KNN), gradient boost (GB), and random forest (RF) were applied to measure accuracy and other parameters. In the second stage, the top three accuracy gained algorithms were chosen, and using voting results were obtained. Pima Indian diabetic (PID) dataset was used and 82% accuracy was obtained from the proposed work. Rawat *et al.* [6] suggested ML algorithms such as NB, SVM, neural network (NN), Adaboost, KNN, and Linear SVM to predict diabetes. NN outperforms others in terms of accuracy. PID dataset was used for the analysis. Ismail *et al.* [7] used 35 ML algorithms such as SVM, decision tree (DT), NB, KNN, logistic regression (LR), RF, artificial neural network (ANN), and multi-layer perceptron (MLP) to predict diabetes. Three datasets retrieved from UCI, MIMIC III, and PIMA diabetes were used Waikato environment for knowledge analysis (WEKA) was used for the implementation. Research by Rajeswari and Ponnusamy [8] discussed SVM and LR to predict diabetes. Dataset was extracted from NC state university. Dataset is split into testing and training data in the ratio of 70% and 30% respectively. 82% accuracy was obtained by SVM for training data and 75% for testing data. Research by Sharma *et al.* [9] discussed supervised ML algorithms: DT, NB, ANN, and LR for the prediction of diabetes. Dataset used was PID which was downloaded from UCI. The experiment was carried out using WEKA 3.8.4. LR performs better than others.

Research by Patra and Khuntia [10] proposed a new classification technique using KNN and standard deviation (SDKNN). In this study, distance calculation was based upon the standard deviation of points. PID dataset was used and acquired from UCI. Dataset was split into 90% and 10% training and testing data respectively. The proposed technique showed an accuracy of 83.2%. Kumari and Bhargavi [11] used SVM, NB, KNN, and DT ML techniques for the early prediction. Dataset of 200 patients taken from health facilities. The results of the experiment showed that the DT outperforms others in terms of accuracy. Kumari *et al.* [12] suggested an ensemble method to predict diabetes. PID dataset was used for the experiment. As base learners

AdaBoost, SVM, LR, RF, NB, Bagging, GB, XGBoost, and CatBoost were used. RF, LR, and NB were used to obtain the final result by using soft voting. 79% accuracy was obtained by this method. According to Khaleel and Bakry [13] diagnose diabetes using KNN, NB, and LR ML algorithms. PID dataset was used for the experiment. Data were split into 70%, 30% testing, and testing respectively. Python was used for the implementation. LR performs better in terms of precision with 94% than others. Research by Joshi and Dhakal [14] predict diabetes using LR and DT. PID dataset was used for the experiment. For the feature selection classification tree was used. After selecting the features, ML algorithms were used. 78.26% accuracy was obtained. Research by Barik *et al.* [15] suggested using RF and XGBoost to predict diabetes. PID dataset was used for the experiment. Python was used for the implementation. The accuracy of XGBoost and RF was 74%, and 71% respectively.

Research by Nagabushanam *et al.* [16] suggested CNN model to predict diabetes. The Pima Indian diabetes database (PIDD) was used for the study. For the purpose of feature extraction Convolutional layer and pooling layers with different relu functions were used. In this, fully connected layers, flatten layers, appropriate dense/output layer and softmax layer were used to classify data. The accuracy of the model was 77.98%. Research by Pethunachiyar [17] suggested that SVM with different kernel functions were applied to predict Diabetes. Dataset was taken from UCI. SVM with a linear kernel had the highest accuracy for diabetes classification. According to Pradhan *et al.* [18] discussed SVM, KNN, NB, LR, AdaBoost, and DT ML algorithms. Dataset was taken from Kaggle. SVM achieved the highest accuracy and F1-score. It also leads to greater sensitivity and recall, as well as a lower Log loss function value. Rajendar *et al.* [19] stated ML algorithms such as DT, LR, RF, and SVM for the prediction of diabetes. The PIDD was used for the study. In comparison to other classifiers, SVM was found to be more accurate in determining the probability of diabetes. According to Tripathi and Kumar [20] discussed four ML algorithms namely linear discriminant analysis (LDA), KNN, SVM, and RF in the predictive analysis of early-stage diabetes. The experimental analysis was performed using PIDD, which was obtained from UCI. RF outperforms other classification algorithms with a maximum accuracy of 87.66%.

2.2. Hybrid ML techniques

These techniques are developed using traditional ML techniques and with different feature extraction techniques such as bio-inspired, metaheuristic, and clustering. Patil *et al.* [21] proposed a hybrid model using mayfly and SVM. Mayfly was used for the optimization of parameters and SVM was for classification. SMOTE statistical technique was also used to balance the cases. Dataset was obtained from UCI and Local hospital for real-time analysis. The anticipated model obtained an accuracy of 94.5%. According to Tan *et al.* [22] anticipated a genetic algorithm (GA)-stacking ensemble learning model for the prediction of Diabetes. GA based on DT was used for feature selection. CNN and SVM were used as base learners and CNN was utilized to obtain the final result. Dataset was taken from Qingdao desensitization physical examination from 1 January 2017 to 31 December 2019. The result of the proposed model was compared with other techniques like GA with KNN, SVM, KNN, LR, CNN, and NB. The proposed algorithm showed an accuracy of 85.08%.

Mallika and Selvamuthukumaran [23] presented a hybrid optimization technique by utilizing the advantages of crow search algorithm (CSA), binary grey wolf optimizer (BGWO) and SVM in the diabetes diagnosis. PID dataset was used and MATLAB was used for the experiment. The accuracy of the proposed technique was 94.8%. Samreen [24] proposed a hybrid algorithm to predict diabetes. In this researcher used ML pipelines for feature selection, feature extraction, and classification. For feature extraction researcher used ANOVA filter, CSA, and singular value decomposition. The classification was performed with several different classifiers like NB, LR, KNN, DT, SVM, RF, AdaBoost, and Gradient Boost as base learners followed by their stacking ensemble. Dataset was acquired from Sylhet Diabetes Hospital in Sylhet, Bangladesh and Python was used for implementation. 98.4% accuracy was obtained.

Azad *et al.* [25] proposed a DM classification model based on SMOTE, GA, and DT (PMSGD. PIDD was used, and it was retrieved from UCI). WEKA was used for the experiment. In terms of CA, CE, accuracy, sensitivity, FM, and AUROC, the proposed system achieved the best results of 82.1256, 17.8744, 0.8070, 0.8598, 0.8326, and 0.8511, respectively. Research by Patil *et al.* [26] suggested a hybrid model having ANN, fuzzy logic, GA, and particle swarm optimization (PSO), to predict diabetes. GA and PSO were applied to optimize parameters for the proposed model. The anticipated model used fuzzification matrix to relate the input patterns with a degree of membership to different classes. PID dataset was downloaded from UCI and MATLAB was used for implantation. Qteat and Awad [27] proposed a hybrid model of PSO and multi-layer perceptron NN (MLPNN) for the classification of diabetes. Dataset was collected from the Palestinian Diabetes Institute DataPal dataset. The result of the proposed model showed an accuracy of 98.73%. MATLAB R2019a was used for the experiment.

Research by Le *et al.* [28] suggested a novel approach to the early prediction of diabetes. In this study, GWO, and an adaptive PSO were used to optimize the MLP and reduce the input attributes. Dataset was obtained from Sylhet Diabetes Hospital of Sylhet, Bangladesh. Results of the proposed approach compared with SVM, DT, KNN, NB, RF, and LR. The projected method achieved an accuracy of 96% for GWO-MLP and 97% for APGWO-MLP. According to Islam *et al.* [29] discussed two new feature selection approaches. For feature extraction, two new approaches based on the fractional derivative and wavelet decomposition were applied. The raw data from the oral glucose tolerance test (OGTT) was pre-processed by using the arithmetical mean to replace missing values. For classification, SVM, NB, RF, AdaBoost, and Bagging models were utilized. Dataset was from a longitudinal clinical study, known as the San Antonio heart study. The proposed ML framework acquired an accuracy of 95.94%. According to Singh and Singh [30] proposed a stacking-based evolutionary ensemble learning system NSGA-II-Stacking for the prediction of type-2 DM. PID dataset was used for the experiment and MATLAB was utilized for implementation. Median values were used to fill the missing values. A multi-objective optimization algorithm was used as the base learner and KNN was used as a meta classifier. The proposed model obtained an accuracy of 83.8%, sensitivity of 96.1%, specificity of 79.9%, F-measure 88.5%, and Roc curve of 85.9%. Table 1 demonstrates ML algorithms used by researchers to detect diabetes. These algorithms will be further used for analysis.

By studying the related work, ML approaches can aid in the early detection of diabetes. These approaches can easily be utilized in hospitals and healthcare institutes. But there are some issues with these studies as well. The related work finds the following research gaps: i) some researchers neglected parameter metrics to show their results. Accuracy is an important factor but other parameters such as relative absolute error (RAE), root mean square error (RMSE), MCC, mean absolute error (MAE), and RRS. are also an important part of performance evaluation, ii) feature selection techniques are not considered by some of the researchers, and iii) time complexity is also neglected by some researchers. The work is divided into multiple sections. Section 1 provides an introduction to diabetes and ML in the medical field and provides information regarding Related work in which research of various researchers is discussed in section 2. Section 3 explains the methodology followed in the paper. In the later section 4, experiment results will be discussed, the last section is conclusion.

Table 1. Related work

ML algorithms	References	Ref.count
SVM	[5]–[8], [11], [12], [17]–[24], [28], [29]	16
DT	[7], [9], [11], [14], [18], [19], [24], [25], [28]	9
RF	[5], [7], [12], [15], [19], [20], [24], [28], [29]	9
KNN	[5]–[7], [10], [11], [13], [18], [20], [22], [24], [28], [30]	12
ANN	[7], [9], [26]	3
NB	[5]–[7], [9], [11]–[13], [18], [22], [24], [28], [29]	12
LR	[7]–[9], [12]–[14], [18], [19], [22], [24], [28]	11
MLP	[7], [27], [28]	3

3. METHOD

This section indicates the steps that are taken to conduct research. From Figure 1 we got a clear view of the steps taken in this work: i) dataset selection, ii) pre-processing, iii) cross-validation, iv) ML algorithms selection, v) prediction, and vi) parameter evaluation. All these steps followed by almost every researcher to conduct research. These steps are further allobrate in the following sub-sections:

3.1. Dataset

For the analysis two similar types of datasets related to diabetes are downloaded (see Table 2). PIDD was retrieved from Kaggle [31]. It has 9 attributes such as preg, plas, pres, skin, insu, mass, pedi, age, and class. 768 instances are present. The first eight attributes are in the feature class while the last attribute is in the target class. Eight attributes have numeric data type while class attribute is of the nominal type.

The dataset on diabetes was taken from the hospital in Frankfurt, Germany, and downloaded from Kaggle [32]. Data has 2001 instances and has 9 attributes such as pregnancies, glucose, blood pressure, SkinThickness, insulin, BMI, DiabetesPedigreeFunction, age and outcome class. 8 attributes are in the feature class whereas the last one is in the target class. The main goal of these datasets was to predict whether a patient is suffering from diabetes or not. For the splitting of data 10 fold cross-validation was used.

3.2. Pre-processing

This phase of method turns less significant knowledge into knowledge that is more pertinent. This process contains some steps, such as data gathering inside a database, relevant data selection, pre-processing of selected data, sampling, and data transformation. Before using the ML algorithm, raw data must first be pre-processed These data may have numerous missing values, numbers outside of the expected range, and noise [33].

Missing data makes it difficult for ML algorithms and approaches to process the input data. Therefore, the information had to be translated into a structural format before any kind of technique could be applied to it. The extraction, transformation, and loading (ETL) process is another name for the data pre-processing stage [34]. In this research work, class balancer filter was applied to maintain the same weight of all instances in the dataset.

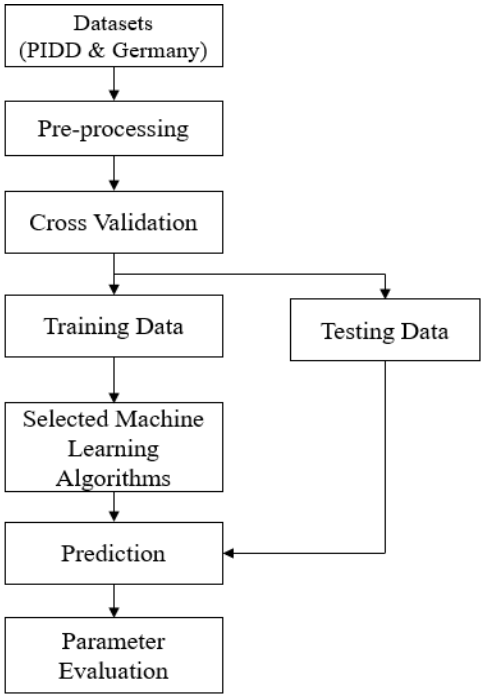


Figure 1. Method for DM

Table 2. Datasets description

Properties	PIDD	Germany
Attribute	9	9
Instances	768	2000
Missing values	No	No
Target class attribute	2 (0, 1)	2(Y,N)

3.3. Algorithm selection

Algorithm selection depends on the dataset and type of prediction. To select the algorithms for the analysis part we have used a variable named Ref. count which counts the number of particular algorithms used in the literature. We have chosen algorithms whose Ref. count >4 from Table 1 and discussed.

NB: the bayes theorem of probability theory is used in the NB algorithm to draw conclusions and classify data based on observed and statistical data. As it is reasonably simple to grasp and highly accurate. It is one of the algorithms that are generally used to produce the most precise predictions relying on a collected dataset. KNN: it evaluates the values of novel data instances by concentrating on 'feature similarity,' which means a value will be allotted to the novel data instances based on how closely it reflects the instances in the training set. It works on the basis of the distance among the data points. Researchers used euclidian equation, and manhattan distance to evaluate the distance among data points.

SVM: it is a classification technique with a specific characterisation of a separating hyperplane. In 2-D space, a hyperplane is a line that splits a plane into two segments, with each class on either side [35]. The supporting vectors are the hyperplane's vectors. Researchers can optimise the distance between hyperplanes. SVM uses a non-linear kernel function to plot data in places where a linear hyperplane is unable to isolate it.

DT: for classification and regression analysis, DT is a widely used non-parametric effective ML algorithm. DT uses predictor data to make continuous, hierarchical decisions regarding the independent variables in order to identify solutions. Instances are classified using DTs by ordering them down the tree from the root to a leaf node. Figure 2 illustrates how an instance is sorted by starting at the tree's root node, evaluating the attribute defined by this node, and then moving along the tree branch based on the attribute's value. The rooted subtree of the new node is then treated in the same way [36].

RF: is an ensemble ML algorithm that entails the creation of numerous DTs using boot-trap aggregation. To put it another way, whenever input is sent to RF, it is routed through each of the DTs. Each tree individually anticipates a classification and votes for the relevant class. The ultimate RF prediction is determined by the majority of votes [37].

LR: The logistic function (LF) is the heart of the algorithm LR. The sigmoid function is another name for the LF. It is an S-shaped arch that can turn any real-valued integer into a number between 0 and 1.

$$\text{Logistic Function} = \frac{1}{1+e^{-\text{value}}} \tag{1}$$

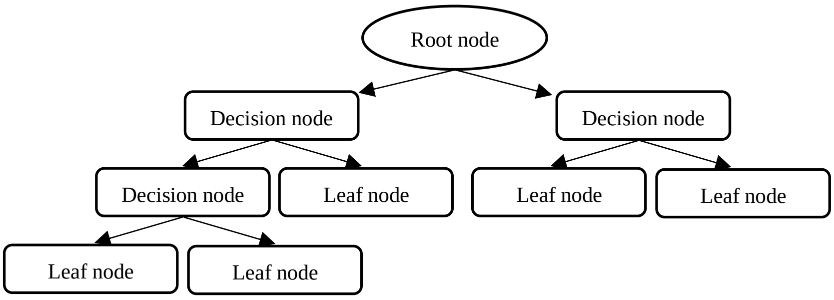


Figure 2. Decision tree architecture

3.4. Software used

WEKA is available as an open software program for ML [38]. The platform facilitates the implementation of many data analysis algorithms and provides a JAVA programming language API, using inbuilt algorithms from a specific application. It has tools for classification, regression, clustering, removing unnecessary characteristics, constructing association rules, and visualizing the dataset. For the experiment WEKA v3.8.6 on AMD Ryzen 5, 5500U with Radeon Graphics with 16 GB RAM on x64 bit Windows 11 operating system is used.

4. EXPERIMENTAL RESULTS AND DISCUSSION

For a comprehensive and unbiased analysis of the algorithms, this paper has two RQs. To respond to RQ1, we explored the wide state-of-the-art in the fields of predictive algorithms and diabetes. To predict diabetes, ML algorithms were selected from Table 1. SVM, DT, RF, KNN, NB, and LR (see Table 1) predictive algorithms were selected for the experiment. To respond to RQ2, we presented a framework for identifying which algorithm performs best for identically structured diabetes datasets. Datasets were taken from Kaggle. There are no missing values in the datasets. In pre-processing phase class balancer filter was applied to maintain the same weight of all instances.

After pre-processing, data was split into testing and training data using 10-fold cross-validation. Selected algorithms from RQ1 were applied to these datasets. Each algorithm was gone through the parameter evaluation phase. The obtained results were shown in Tables 3 and 4.

$$Accuracy = \frac{(TP+TN)}{(\text{TP+FP+FN+TN})} \tag{2}$$

$$Recall \ = \frac{TP}{TP+FP} \tag{3}$$

$$Precision \ = \frac{TP}{TP+FN} \tag{4}$$

For the analysis, parameter metrics and error rates used are: accuracy, precision, recall, ROC area, kappa value, MAE, RMSE, RAE, and RRSE. Where means TP: true positive, the number of cases correctly identified as patient. TN: true negative, the number of cases incorrectly identified as patient. FP: false positive, the number of cases incorrectly identified as healthy. FN: false negative, the number of cases incorrectly identified as healthy

Table 3. PID dataset analysis

	Accuracy (%)	Precision	Recall	ROC area	MCC	Kappa value
NB	72.6	0.72	0.72	0.81	0.45	0.45
KNN	66.1	0.67	0.66	0.65	0.30	0.32
SVM	74.3	0.74	0.74	0.74	0.48	0.48
DT	71.8	0.71	0.71	0.71	0.43	0.43
RF	64.9	0.65	0.65	0.64	0.47	0.29
LR	74.0	0.74	0.74	0.83	0.47	0.48

Table 4. Germany dataset analysis

	Accuracy (%)	Precision	Recall	ROC area	MCC	Kappa value
NB	76.5	0.75	0.76	0.82	0.47	0.45
KNN	98.7	0.98	0.98	0.97	0.97	0.97
SVM	77.0	0.76	0.77	0.71	0.50	0.45
DT	94.5	0.94	0.94	0.97	0.87	0.87
RF	98.7	0.98	0.98	0.99	0.97	0.98
LR	77.6	0.77	0.77	0.83	0.50	0.47

Kappa value: the kappa statistic measures how well the instances are classified by the ML classifier and classify the labelled data as ground truth while controlling for the anticipated accuracy of a random classifier. It examines how effective a classifier is for a specific dataset.

$$Kappa = \frac{i_0 - i_e}{1 - i_e} \quad (5)$$

Where i_0 is overall accuracy and i_e is a measure of the agreement between the model predictions and the actual class values as if happening by chance.

ROC curve: is calculated by contrasting the true positive rate (TPR) against the false positive rate (FPR) at various threshold levels and efficiently divides the signal from the noise. Mean absolute error: it reflects the gap between the original and predicted values as determined by averaging the absolute difference across the data set [39]. RMSE: it is a prominent approach to evaluating the error in the model for predicting statistical data. RMSE scores between 0.0 and 0.5 which implies that the model can accurately predict the data. RAE: it is a method of evaluating the effectiveness of a predictive model. It is expressed as a ratio, contrasting mean errors to trivial errors [40]. Root relative squared error (RRSE): it is a basic indicator that provides an idea of how well a model performs. Furthermore, it is a variation of the relative squared error (RSE). Matthews correlation coefficient: it examines categorization quality by accounting for true and false positives and negatives. In this 1 represents a perfect prediction, 0 reflects no better than a random prediction, and -1 indicates an absolute conflict between prediction and observation [41].

The study discovered that SVM exceeds others in terms of accuracy with 74.3% for the PID dataset (Table 3). NB scored 72.6%, KNN 66.1%, DT 71.8%, RF 64.9% and LR scored 74.0% in terms of accuracy. After SVM, LR performs better with the nearly same accuracy of 74.0% (Figure 3). So, we can say that for the PIDD SVM and LR are better choices in terms of accuracy.

The results of the research for the second dataset revealed that KNN and RF exceed others in terms of accuracy, with a score of 98.7%. (Table 4). NB scored 76.5%, SVM scored 77.0, DT 94.5% and LR scored 77.6% in terms of accuracy. After KNN and RF, DT showed an accuracy of 94.5% (Figure 4). Consequently, KNN and RF are the best-suited choices for this dataset. MCC is used to (in Tables 3 and 4) check the correlation between the true and predicted values. If the correlation is higher, prediction results will also be high. For PID Dataset, the KNN MCC score is low whereas for the Germany dataset MCC score for each classifier is moderately good.

Tables 5 and 6 discussed about the Error rates for each classifier. The error rate should be minimum for optimal results. RAE ranges between 0 to1. 0 implying that the classifier fitted well for the given dataset whereas 1 implies a poor classifier. RRSE (in Tables 5 and 6) values should be low for a better prediction from a classifier. In the PID dataset RRSE values are very high as compared to Germany dataset. For the Germany dataset, KNN and RF generate low RRSE values and hence can be considered as good classifiers. Time consumption (in Tables 5 and 6) for each classifier is less than 1 which is a good indication.

By analyzing the ROC areas (Figure 5) of both datasets it is found that LR for PIDD and RF for the German dataset perform better (from Tables 3 and 4). If kappa values (Figure 6) of both datasets are analysed, for the PID dataset selected classifiers are not enough strong but for the Germany dataset, some classifiers like RF, DT, and KNN are enough strong. In the analysis part accuracy, precision, recall, F-measure, MCC, and ROC area are used to examine the reliability, whereas MAE, RMSE, RAE, and RRSE (in Tables 5 and 6) are used to examine the error rates of the particular classifier.

Overall, we can infer from the results of the two datasets that LR can be preferred for classification by looking at accuracy and the ROC curve. As accuracy determines how precisely a dataset is being classified by a particular classifier. So, LR can perform well for both datasets.

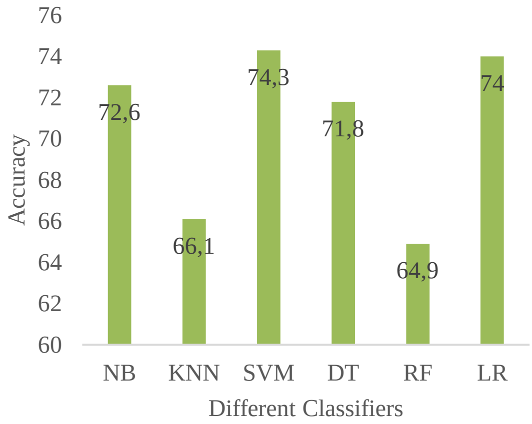


Figure 3. Accuracy of different algorithms (PIDD)

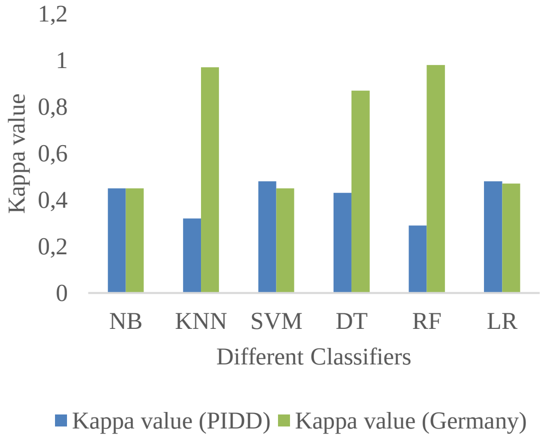
Figure 4. Accuracy of different algorithms (Germany)

Table 5. PID dataset error rate analysis

	MAE	RMSE	RAE	RRSE%	Time (sec)
NB	0.30	0.42	0.61	85.7	0.01
KNN	0.33	0.58	0.67	116.1	0.00
SVM	0.25	0.50	0.51	101.3	0.09
DT	0.34	0.47	0.68	95.9	0.08
RF	0.33	0.41	0.66	83.9	0.46
LR	0.33	0.40	0.66	81.9	0.08

Table 6. Germany dataset error rate analysis

	MAE	RMSE	RAE	RRSE%	Time (sec)
NB	0.30	0.42	0.61	85.2	0.00
KNN	0.013	0.11	0.02	22.4	0.00
SVM	0.24	0.49	0.49	99.7	0.08
DT	0.07	0.23	0.14	46.7	0.06
RF	0.07	0.13	0.14	26.7	0.52
LR	0.33	0.40	0.67	81.7	0.02



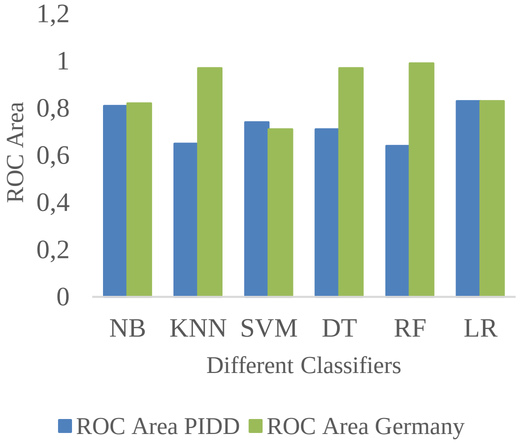


Figure 5. ROC area of both datasets

Figure 6. Kappa values of both datasets

5. CONCLUSION

Diabetes is a widespread disease that affects the majority of the world's population. Diabetes must be detected early because it can lead to other complications. For the automated detection of diabetes predictive algorithms are the better choices. The following ML predictive algorithms are presented in this study to aid in diabetes prediction: NB, KNN, SVM, DT, RF, and LR. They were chosen on the basis of current research work on diabetes and predictive algorithms. The experiment was piloted using the PID and Germany diabetes datasets and implemented using WEKA software. It can be concluded from the experiment that SVM and LR outperform in terms of accuracy for PIDD and for ROC area LR performs better. For Second, we can conclude that KNN and RF works better in terms of accuracy. In terms of the ROC area, RF outperforms. This research also discussed about the error rates for the particular classifiers. The overall analysis of the research infers that LR can be preferred for both datasets. This research may aid healthcare institutions in the early detection of diabetes, saving doctors time and effort while also being cost-effective for patients. The study mentioned some of the hybrid models but we did not experiment related to them. So, in the near future, we will use hybrid models for the experiment and analyse their performance against these datasets as well as with some real-time datasets.

REFERENCES

[1] P. Prabhu and S. Selvabharathi, "Deep belief neural network model for prediction of diabetes mellitus," in *2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC)*, 2019, pp. 138–142, doi: 10.1109/ICISPC.2019.8935838.

[2] N. A. Farooqui, . R., and A. Tyagi, "Prediction model for diabetes mellitus using machine learning techniques," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 3, pp. 292–296, 2018, doi: 10.26438/ijcse/v6i3.292296.

[3] TFPR editorial, "Diabetes is a pandemic in India. But the Sugar Association wants people to consume more!," *The Future of Public Relations*, 2020.

[4] "National Diabetes and Diabetic Retinopathy Survey - INSIGHTSIAS." https://www.insightsonindia.com/2019/10/11/national-diabetes-and-diabetic-retinopathy-survey/ (accessed Dec. 08, 2022).

[5] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, "Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques," *Hindawi*, vol. 2022, no. Special Issue, 2022, doi: 10.1155/2022/6521532.

[6] V. Rawat, S. Joshi, S. Gupta, D. P. Singh, and N. Singh, "Machine learning algorithms for early diagnosis of diabetes mellitus: A comparative study," *Mater. Today Proc.*, vol. 56, part 1, pp. 502–506, 2022, doi: 10.1016/j.matpr.2022.02.172.

[7] L. Ismail, H. Materwala, M. Tayefi, P. Ngo, and A. P. Karduck, "Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation," *Arch. Comput. Methods Eng.*, vol. 29, no. 1, pp. 313–333, 2022, doi: 10.1007/s11831-021-09582-x.

[8] S. V. K. R. Rajeswari and V. Ponnusamy, "Prediction of diabetes mellitus using machine learning," *Ann. Rom. Soc. Cell Biol.*, vol. 25, no. 5, pp. 17–20, 2021.

[9] A. Sharma, K. Guleria, and N. Goyal, "Prediction of Diabetes Disease using Machine Learning Model," in *Lecture Notes in Electrical Engineering (2021) 733 LNEE 683-692*, 2021, no. March, doi: 10.1007/978-981-33-4909-4.

[10] R. Patra and B. Khuntia, "Analysis and prediction of pima Indian diabetes dataset using SDKNN classifier technique," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1070, no. 1, pp. 1–14, 2021, doi: 10.1088/1757-899X/1070/1/012059.

[11] K. S. Kumari and K. Bhargavi, "Performance Analysis of Diabetes Mellitus Using Machine Learning Techniques," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 6, pp. 225–230, 2021.

[12] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *Int. J. Cogn. Comput. Eng.*, vol. 2, no. November 2020, pp. 40–46, 2021, doi: 10.1016/j.ijcce.2021.01.001.

[13] F. Alaa and A. M. Al-bakry, "Diagnosis of diabetes using machine learning algorithms," *Mater. Today Proc.*, 2021, doi: 10.1016/j.matpr.2021.07.196.

[14] R. D. Joshi and C. K. Dhakal, "Predicting Type 2 Diabetes Using Logistic Regression andMachine Learning Approaches," *Int. J. Environ. Res. Public Health*, vol. 18, no. 14, p. 7346, 2021, doi: 10.3390/ijerph18147346.

[15] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques," *Smart Innov. Syst. Technol.*, vol. 153, no. January, pp. 399–409, 2021, doi: 10.1007/978-981-15-6202-0_41.

[16] P. Nagabushanam, N. C. Jayan, C. Antony Joel, and S. Radha, "CNN architecture for diabetes classification," *2021 3rd Int. Conf. Signal Process. Commun. ICPSC 2021*, no. May, pp. 166–170, 2021, doi: 10.1109/ICSPC51351.2021.9451724.

[17] G. A. Pethunachiyar, "Classification Of Diabetes Patients Using Kernel Based Support Vector Machines," in *2020 International Conference on Computer Communication and Informatics, ICCCI 2020*, 2020, pp. 22–25, doi: 10.1109/ICCCI48352.2020.9104185.

[18] R. Pradhan, M. Aggarwal, D. Maheshwari, A. Chaturvedi, and D. Sharma, Kumar, "Diabetes Mellitus Prediction and Classifier Comparitive Study," in *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control*, 2020, pp. 133–139, doi: 10.1109/PARC49193.2020.236572.

[19] S. Rajendar, R. Thangaraj, J. Palanisamy, and V. K. Kaliappan, "Comparative analysis of classifier models for the early prediction of type 2 diabetes," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 7, pp. 2184–2194, 2020.

[20] G. Tripathi and R. Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning," in *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (2020) 1009-1014*, 2020, pp. 1009–1014, doi: 10.1109/ICRITO48877.2020.9197832.

[21] R. Patil, S. Tamane, S. A. Rawandale, and K. Patil, "A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 1, pp. 524–533, 2022, doi: 10.11591/ijece.v12i1.pp524-533.

[22] Y. Tan, H. Chen, J. Zhang, R. Tang, and P. Liu, "Early Risk Prediction of Diabetes Based on GA-Stacking," *Appl. Sci.*, vol. 12, no. 2, p. 632, 2022, doi: 10.3390/app12020632.

[23] C. Mallika and S. Selvamuthukumaran, "A Hybrid Crow Search and Grey Wolf Optimization Technique for Enhanced Medical Data Classification in Diabetes Diagnosis System," *Int. J. Comput. Intell. Syst.*, vol. 14, no. 1, 2021, doi: 10.1007/s44196-021-

00013-0.
[24] S. Samreen, "Memory-efficient, accurate and early diagnosis of diabetes through a machine learning pipeline employing crow search-based feature engineering and a stacking ensemble," *IEEE Access*, vol. 9, pp. 134335–134354, 2021, doi: 10.1109/ACCESS.2021.3116383.
[25] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimed. Syst.*, vol. 28, no. 4, pp. 1289–1307, 2022, doi: 10.1007/s00530-021-00817-2.
[26] R. Patil, T. Sharvari, and R. Nirmal, "Hybrid ANFIS-GA and ANFIS-PSO Based Models for Prediction of Type 2 Diabetes Mellitus," in *Advances in Intelligent Systems and Computing*, vol. 1227, 2021, pp. 11–23.
[27] H. Qteat and M. Awad, "Using Hybrid Model of Particle Swarm Optimization and Multi-Layer Perceptron Neural Networks for Classification of Diabetes," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 3, pp. 11–22, 2021, doi: 10.22266/ijies2021.0630.02.
[28] T. M. Le, T. M. Vo, T. N. Pham, and S. V. T. Dao, "A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic," *IEEE Access*, vol. 9, pp. 7869–7884, 2021, doi: 10.1109/ACCESS.2020.3047942.
[29] M. S. Islam, M. K. Qaraqe, S. B. Belhaouari, and M. A. Abdul-Ghani, "Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes," *IEEE Access*, vol. 8, pp. 120537–120547, 2020, doi: 10.1109/ACCESS.2020.3005540.
[30] N. Singh and P. Singh, "Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus," *Biocybern. Biomed. Eng.*, vol. 40, no. 1, pp. 1–22, 2020, doi: 10.1016/j.bbe.2019.10.001.
[31] "Pima Indians Diabetes Database | Kaggle." https://www.kaggle.com/uciml/pima-indians-diabetes-database (accessed Jul. 29, 2021).
[32] "diabetes | Kaggle." https://www.kaggle.com/datasets/johndasilva/diabetes?resource=download (accessed Apr. 23, 2022).
[33] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *Int. J. Comput. Sci.*, vol. 1, no. 2, pp. 111–117, 2006, doi: 10.1080/02331931003692557.
[34] R. Kimball and J. Caserta, *The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data*. New Jersey, USA: John Wiley & Sons, 2015.
[35] G. Mutlu, "SVM-SMO-SGD : A hybrid-SVM-SMO-SGD: A hybrid-parallel support vector machine algorithm using sequential minimal optimization with stochastic gradient descent," *Parallel Comput.*, vol. 113, no. July, 2022, doi: 10.1016/j.parco.2022.102955.
[36] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," *Int. J. Comput. Sci.*, vol. 9, no. December 2013, pp. 272–278, 2012.
[37] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomed. Signal Process. Control*, vol. 52, pp. 456–462, 2019, doi: 10.1016/j.bspc.2017.01.012.
[38] Waikato, "Weka 3: machine learning software in Java," *GitHub*, 2021.
[39] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.
[40] R. Naseem *et al.*, "Empirical assessment of machine learning techniques for software requirements risk prediction," *Electron.*, vol. 10, no. 2, pp. 1–19, 2021, doi: 10.3390/electronics10020168.
[41] L. Ali *et al.*, "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," *IEEE Access*, vol. 7, pp. 54007–54014, 2019, doi: 10.1109/ACCESS.2019.2909969.

BIOGRAPHIES OF AUTHORS

Ms. Kirti Kangra is pursuing her Ph.D. from Guru Jambheshwar University of Science and Technology, Hisar, Haryana in Computer Science and Engineering. She has teaching experience for 2 years. She has completed her B.Tech in Information and Technology from Vaish College of Engineering, Rohtak, and M.Tech in Computer Science and Engineering from the University Institute of Engineering and Technology affiliated with Maharishi Dayanand University, Rohtak, Haryana. Her areas of research include machine learning, data mining, and artificial intelligence. She can be contacted at email: Kirtikangra98@gmail.com.

Dr. Jaswinder Singh is working as an Associate Professor in the Department of Computer Science and Engineering at Guru Jambheshwar University of Science and Technology, Hisar, Haryana. He has teaching experience of more than 18 years and he has published more than 25 research papers in international journals and conferences. He has completed his Ph.D. in Computer Science and Engineering from Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Sonepat, Haryana, and completed his M.Tech in Computer Science and Engineering from Kurukshetra University, Kurukshetra, Haryana. His areas of research include machine learning, opinion mining, web information retrieval, search engine optimization, web mining, information processing, information system, and social network analysis. He can be contacted at email: jaswinder_singh@gjust.org.