





Liabile Characteristics Measure and Anticipate the Diabetes Disease Using Machine Learning Tools

M. Murad Hossain¹, Md. Rana Ahmed¹, M. Zahid Hasan^{1,*}, M. Sultana², K. Fatema¹

¹Department of Statistics, Faculty of Science, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj-8100, Bangladesh

sadid.shimul@gmail.com, ranastat1994@gmail.com, zahid1680@gmail.com, khairunnasafatema@gmail.com

²Department of Mathematics, Faculty of Science, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj-8100, Bangladesh

s.mitu@bsmrstu.edu.bd

*Correspondence: zahid1680@gmail.com

ABSTRACT. Diabetes is a cardiovascular disease. It is not only an epidemic in Bangladesh but also in the whole world that is increasing rapidly. At an early period of human life, machine learning techniques are used to predict diabetes datasets. In our research paper, we use the Pima diabetes dataset from the Kaggle UCI machine learning data repository. For diabetic patients and doctors, machine learning techniques are both cost-effective and time-saving. We apply KNN, Nave Bayes, Random forest, Support vector machine, Simple logistic, and J48 to Pima datasets. Besides these algorithms, we may develop an ensemble (Vote) hybrid model with WEKA software by combining individual methods that provide the best performance and accuracy. Also, try to make a comparison among all machine learning tool's accuracy and performance with the proposed ensemble model.

1. Introduction

Diabetes is a cardiovascular disease which is the result of high amount of glucose in the blood. We obtain this blood glucose from food and this is the main source of energy [1]. Rather than being a clinical or diagnostic category, high blood glucose is seen as a statistical phenomenon [2]. Insulin is one kind of hormone that is made by the pancreas and it helps glucose from food to get into the cells to create energy. Sometimes the human body can't make enough or any insulin or doesn't use insulin well. As a result, Glucose stays in our blood without reaching to the cells and over time, these excessive amount of

Received: 4 Jun 2022.

Key words and phrases. diabetes; machine learning; accuracy; roc-curve; performance; hybrid model.

glucose in the blood causes health problems [1]. A recent meta-analysis showed that in Bangladesh, the prevalence of diabetes among adults had increased substantially, from 4% in 1995 to 2000 and 5% in 2001 to 2005 to 9% in 2006 to 2010 and the International Diabetes Federation anticipated that the prevalence will be 13% by 2030, [3]. The Federation also claimed that in our country, 6.9 million people are suffering from diabetes and it will become 13.7 million by 2045 [4]. The Bangladesh Bureau of Statistics (BBS) estimates that there are 6 million diabetic patients in the country with an increase of around 180,000 new patients each year [5]. The patients may experience life-threatening complications like heart attack, stroke, damages in kidney, failure of nervous system and blindness. In a survey conducted by the Bangabandhu Sheikh Mujib Medical University (BSMMU) on 2000 adults in Dhaka slums in 2016, it has been found that 19 percent of adults had diabetes [6]. At the age of 35 or more, one in three people are diabetic or pre-diabetic where only 12% of them have their condition under control [7]. Another survey shows, in 2015 one in 11 adults has diabetes and in 2040 one in 10 adults will have diabetes in Bangladesh [8]. The BIRDEM hospital diagnosed 115 new diabetic children in 2010 and four percent of them were suffering from type 2 diabetes. In 2015, the number of newly diagnosed children rose to 419, of which 13 percent were found with type 2 diabetes [9]. In middle- and low-income countries, Diabetes prevalence has been rising more rapidly [10]. In Asia this prevalence is increasing rapidly, such as in 2017 according to the Index Mundi Saudi Arabia diabetes prevalence rate is 17.72%, in United Arab Emirates is 17.26% and so on [11]. Machine learning uses artificial intelligence (AI) which provides the systems the ability to learn automatically and improve from experience without any explicit programme. Data mining involves methods regarding machine learning, statistics and database systems in order to discover patterns in large datasets. So data mining is the application and machine learning is the algorithm that we use. Using machine learning algorithm is the process of data mining. Since Bangladesh is the country of the least developing. The costs of treating and managing diabetes in developing countries are limited. So we use machine learning algorithm to classify diabetes and which classification algorithm is better for the diabetes according to its result performance.

2. Literature Review

Machine learning techniques (MLT) can be applied in the prediction of the medical datasets. For instance, a total of 768 observations, a data set from PIDD (Pima Indian Diabetes Data Set). Using these algorithms, we can propose a hybrid model with the combination of individual techniques or methods, in order to develop the performance and accuracy [12]. Data mining is the process which is mainly used for the selection of special characteristics, information extraction and discovering the unknown pattern and relations from the unstructured data. This research work provides the description of chosen classification models and datasets, an evaluation and a comparison of the performance of 5 classification techniques based on the chosen dataset. This work demonstrates the decision tree as the best technique for the prediction purpose in diabetic patients [13]. To analyse the performance of health care data and disease prediction, different data mining tools are being compared. In a paper, Naïve Bayes classification algorithm has been used in data mining tools Orange, Weka, Rapidminer where Weka provides the best accuracy [14]. After the selection of feature and unbalanced process, the diabetes follow-up data of the New Urban Area of Urumqi, Xinjiang, were used as input variables and the experimental results show that Adaboost algorithm produces better classification results [15]. For the Pima Indian Diabetes Data Set, the proposed ensemble method (PEM) provides high accuracy of 90.36 [16]. Applying different procedures, the diabetic data is being forecasted and it has been found that the Naïve Bayes and C4.5 algorithm system perform better with satisfactory results [17]. That paper presents the analysis of four classification algorithms namely J48, Random tree, Decision tree and Naive Bayes for Diabetic dataset. The experimental result shows that J48 provides better accuracy than the Random tree, Decision tree and Naive Bayes [18]. Using Machine learning algorithms, a precision value equal to 0.770 has been obtained and a recall equal to 0.775 using the Hoeffding Tree algorithm [19]. Participants from two communities in Guangzhou, China; 735 patients confirmed to have diabetes or pre-diabetes and 752 were found to be the normal controls. It has been seen that the decision tree model (C5.0) had the best classification accuracy and the ANN had the lowest one [20]. There are 8 features in the Pima Indians Diabetes data set after removing F3 (Diastolic Blood Pressure) and F7 (Diabetes Pedigree

Function) features from the data set and the 81.89% accuracy of the model proves that their proposed model performs better with less number of features [21]. In that research work, it has been seen that the classification techniques like J48, CART, SVMs, and KNN were frequently applied on the medical dataset to find the optimal solution for Diabetes. The results support the superior performance of J48 technique significantly than the three other techniques for the classification of diabetes data [22]. Several studies have been carried out regarding insulin-dependent and adult-onset diabetes. Most of the studies that carried out, deals with classification aspects. While dealing with classification problem, Naïve Bayes algorithm has proved to be the most efficient one in many studies. Hence, this method has been applied in many work [23]. According to the results of Adaptive Network-based Fuzzy Inference System (ANFIS) and Rough Set methods, ANFIS has proved to be more successful and reliable method for diabetes drug planning objective [24]. The Cart provide the highest performance accuracy which is 83.2% for the diagnosis of diabetes[25]. In that paper, they review data mining techniques in health care management that predicts the disease, diagnosis diabetes so that the health care management can alert the human being regarding diabetes based upon that prediction [26]. The data mining tool WEKA has been used as an API of MATLAB for generating the modified J-48 classifiers which has an accuracy rate of 99.87% while others can show a maximum of 77.21% accuracy [27]. For the diagnosis purpose of type II diabetes, the results of a study revealed that Naive Bayes (having accuracy rate of 76.95%) showed the highest accuracy level [28]. SVM that contains the Radial basis function kernel is used for classification purpose. The different performance parameters of the SVM and RBF such as the classification accuracy, sensitivity, and specificity ; have found to be high and that's why it can be considered to be a good option for the classification process [29]. These artificial intelligence algorithms are cost effective and time saving for diabetic patients as well as doctors. In a research paper, Diabetes Mellitus was detected using K- Nearest neighbor algorithm which is one of the most important techniques of A.I [5]. A Random forest is defined as the combination of decision trees which can help in predicting data accurately. The Redistribution error rate in case of random forest is less than that of decision tree [30]. These methods have been tested with data samples downloaded from UCI machine learning data repository

where the KNN ($k=1$) and Random Forest performed much better than the other three classifiers and they provide 100% accuracy [31]. A study that applies the Adaboost and bagging ensemble techniques using J48 (c4.5) decision tree as a base learner along with standalone data mining technique J48 to classify patients with diabetes mellitus using diabetes risk factors and the result demonstrates that the overall performance of the Adaboost ensemble method is much better than bagging and standalone J48 decision tree [32]. Our research is different and essential from other research because the most of the researcher use Pima diabetes data. They were using either Base level classifier or Meta level classifier; in our research paper firstly we use a base-level classifier then create a Meta level classifier combining with the base level classifier which provides better accuracy than the base level classifier. Some of them use data mining tools Orange, Rapidminer, Knime, Oracle, Teradata, MATLAB etc. But we know for the diabetes datasets WEKA software provides the best accuracy [33]. However, articles summarizing the various approaches and effectiveness of are rarely used the machine learning techniques in addition as a hybrid model to predict the diabetes patients and does not show any comparison with the proposed ensemble model. Through this study, we try to predict Diabetes patient's people through the proposed ensemble model and other machine learning algorithm. We also try to measure the performance of various machine learning tools within the perspective of validation and accuracy. Here, we use WEKA software version 3.8.3 in our research. For diabetes dataset Meta soft vote classifier with average probability provides high accuracy.

3. Methodology

3.1 Data collection and Data attributes

The Indian Pima diabetes dataset is collected from Kaggle UCI Machine Learning repository, which constraints 768 observations with corresponding 9 attributes, which are- the number of times pregnant, plasma glucose concentration in an oral glucose tolerance test, diastolic blood pressure (in mm Hg), triceps skin fold thickness (in mm), 2-Hour serum insulin(in μ U/ml), body mass index(in weight in kg/(height in m)²), diabetes pedigree

function, age(in years), class (Diabetes, Non-diabetes). For our analysis, we use WEKA version 3.8.3 software.

3.2 Machine Learning Classifier

There are many machine learning classifier which is based on predefined value. Machine learning classifiers use trained data with stratified cross-validation data and apply in the test data. In our research, we use Machine Learning Classifier algorithm these are Random Forest (RF) algorithm, J48 algorithm, Support Vector Machine (SVM) algorithm, K Nearest Neighbors (KNN) algorithm, Naïve Bayes (NB) algorithm, Bayes Nets (BN) algorithm, Simple Logistic Regression (SLR) algorithm and Meta level algorithm (Vote).

3.3 Random Forest Algorithm

Random Forest is the tree base machine learning algorithm. Random forest is the development form bagging machine learning. The concept of Random forests represents the combination of tree predictors where each tree depends on the values of a random vector that is sampled independently and has the same distribution for all trees in the forest [34].

3.4 Support Vector Machine Algorithm

Support vector machine algorithm is also known as binary approach algorithm because it is used for binary classification like present or absence, on or off, normal or abnormal, impactful or none impact-full. In this study, it has been used for the prediction of diabetes: that is diabetes or non-diabetic. It is used by maximizing the margin of the distance between the variables in the hyper plane. It is used for both regression and classification purposes. For classification purpose, we find a hyper plane that provides the best partition of the different classes (maximizes the distance between the data and the nearest data point in each class). This hyper-plane can be either a straight line (linear) or any curve. In order to perform non-linear classification, SVM uses Kernel trick. Kernel can be used to convert the low dimensional space into the high dimensional space. In our method, we use a polynomial kernel.

3.5 J48 Algorithm

J48 decision tree classification is the process of building a model of classes from a set of records which contain class labels. This is an extension of ID3 and considered to be an open-source Java implementation of the C4.5 algorithm.

3.6 K-Nearest Neighbors algorithm

K-Nearest Neighbour algorithm is one of the classification algorithms. This technique uses similarity measures to classify the new belongings. The values of k always take positive integer numbers. According to this algorithm, the training data are stored and based on the neighbors or nearest prediction of test data is complete.

3.7 Naïve Bayes Algorithm

The Bayesian classification is a supervised learning method and it is also known as statistical classification, which depends on the frequency table of prior information and posterior information according to the conditional distribution.

3.8 Simple Logistic algorithm

Logistic regression is a statistical method that analyses a dataset containing one or more independent variables that determine an outcome. Here a dichotomous variable (in which there are only two possible outcomes) measure the outcome. That is, in case of logistic regression, the dependent variable is binary or dichotomous.

3.10 Ensemble Vote Classifier algorithm

For prediction purposes, individual prediction algorithms do not provide better and efficient results. So, in order to increase the accuracy and better performance, it is required to predict those individual prediction algorithms into one by applying the combination of the prediction of a single classifier. This collaborative approach is helpful in solving the limitation of distinct classifiers by combining it into one and providing better accuracy. The Ensemble Vote Classifier is a one kind of meta-classifier which combines the similar or conceptually different machine learning classifiers for classification for soft voting with

average probability. In this situation, the averaging probability is free from bias than the majority probability. The weight function is

$$\hat{y} = \arg \max_i \sum_{j=1}^m w_j P_{ij}$$

where, w_i, w_j is the weight that can be assigned to the ij^{th} classifier.

3.11 Collaborative (Ensemble) model

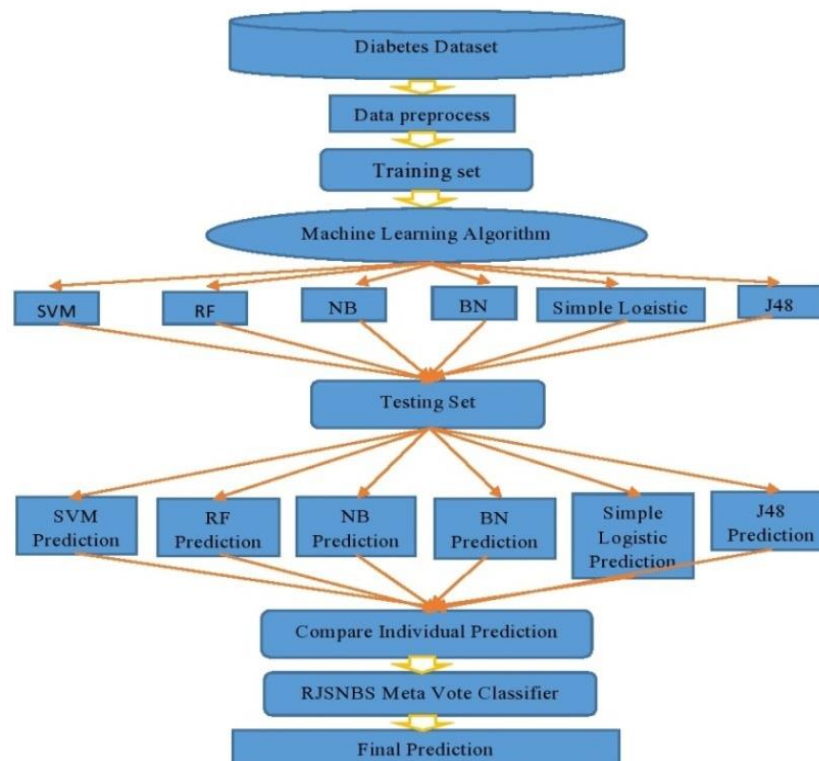


Figure 1. Steps of Study frame.

4. Results and Analysis

This section contains the detail analysis and prediction regarding diabetes diseases applying the Machine Learning algorithm. We divided this section into three parts. The first part represents “Base level classifier” Machine learning algorithm performance on diabetes datasets. The second part represents “Meta level classifier” Machine learning algorithm on diabetes datasets. At last comparison of Base level classifier and Meta level classification Machine learning algorithm based on error rate, recall, precision, recall, F-

measure, ROC area curve and accuracy. The performance evaluation is based on the machine learning algorithm according to the diabetes data. The Base level and Meta level classification performance are given below:

Table 1. Different Characteristics and its value for all Machine learning Technique.

Characteristics	RF	SVM	J48	KNN	NB	LR	EVC
Mean absolute error	0.3124	0.2292	0.316	0.2988	0.2843	0.3171	0.2930
Root mean square error	0.4051	0.4787	0.4463	0.5453	0.4168	0.396	0.3983
Kappa statistic	0.461	0.4631	0.4164	0.3304	0.4664	0.4827	0.4871
Percentage of correctly classified	75.9115	77.0833	73.8281	70.1823	76.3021	77.7344	77.8646
Percentage of incorrectly classified	24.0885	22.9167	26.1719	29.8177	23.6979	22.2656	22.1354

Table 2. Details of accuracy by class for different machine learning techniques.

Method	TP rate	FP rate	Precision	Recall	ROC area	F-measure	Accuracy %	Class
RF	0.619	0.166	0.667	0.619	0.818	0.642	75.9115	Diabetes
	0.834	0.381	0.803	0.834	0.818	0.818		Non-diabetes
	0.759	0.306	0.756	0.759	0.818	0.757		Weighted Average
SVM	0.541	0.106	0.732	0.541	0.718	0.622	77.0833	Diabetes
	0.894	0.459	0.784	0.894	0.718	0.836		Non-diabetes
	0.771	0.336	0.766	0.771	0.718	0.761		Weighted Average
J48	0.597	0.186	0.632	0.597	0.751	0.614	73.8281	Diabetes
	0.814	0.403	0.790	0.814	0.751	0.802		Non-diabetes

KNN	0.738	0.327	0.735	0.738	0.751	0.736	70.1823	Weighted Average
	0.530	0.206	0.580	0.530	0.650	0.554		Diabetes
	0.794	0.470	0.759	0.794	0.650	0.776		Non- diabetes
NB	0.702	0.378	0.696	0.702	0.650	0.698	76.3021	Weighted Average
	0.612	0.156	0.678	0.612	0.818	0.643		Diabetes
	0.844	0.388	0.802	0.844	0.818	0.823		Non- diabetes
LR	0.763	0.307	0.759	0.763	0.818	0.760	77.7344	Weighted Average
	0.567	0.110	0.734	0.567	0.832	0.640		Diabetes
	0.890	0.433	0.793	0.890	0.832	0.839		Non- diabetes
VMC	0.777	0.320	0.773	0.777	0.832	0.832	77.8646	Weighted Average
	0.575	0.112	0.733	0.575	0.833	0.644		Diabetes
	0.888	0.425	0.796	0.888	0.833	0.839		Non- diabetes
	0.779	0.316	0.774	0.779	0.833	0.771		Weighted Average

The prediction and classification of diabetes and non-diabetes with Random Forest algorithm we see that Precision is 0.667 and 0.803, Recall 0.619, ROC curve area 0.818 and 0.834, Accuracy 75.9115% for both diabetes and non-diabetes. For the Support vector machine algorithm, we see that from both the above table and figure Precision is 0.732 and 0.784, Recall 0.541 and 0.894, ROC Area 0.7175 and Accuracy 77.0833% for both diabetes and non-diabetes. Whereas J48 algorithm shows that Precision is 0.632 and 0.790, Recall 0.597 and 0.814, ROC Area 0.7512 and Accuracy 73.8281% for both diabetes and non-diabetes. Then for the KNN algorithm shows Precision is 0.580 and 0.759, Recall 0.530

and 0.794, ROC area 0.650 and Accuracy 70.1823% for both diabetes and non-diabetes. The prediction and classification of diabetes and non-diabetes with Naïve Bayes algorithm we see that from both above table and figure Precision is 0.678 and 0.802, Recall 0.612 and 0.844, ROC area 0.818 and Accuracy 76.3021% for both diabetes and non-diabetes. Simple Logistic algorithm depicts that Precision is 0.734 and 0.793, Recall 0.567 and 0.890, ROC Area 0.832 and Accuracy 77.7344% for both diabetes and non-diabetes. Meta RJSNBS classifier algorithm reveals that Precision is 0.733 and 0.796, Recall 0.575 and 0.888, ROC area 0.833 Accuracy 77.8646% for both diabetes and non-diabetes.

4.1 Comparison of different Machine Learning Classifier algorithms

In our research paper basically, we focus on the Machine Learning algorithm performance based on its true positive rate, false-positive rate, ROC area, F-measure recall, precision, absolute error rate, root mean square error rate, percentage of correctly classified and percentage of incorrectly classified. So our main purpose is to find out the best Machine Learning algorithm which is the best to correctly classify the diabetes data set according to the predefined values. After individual representation, we now compare nine Machine Learning algorithms in the same frame by graphical representation of accuracy and ROC curve.

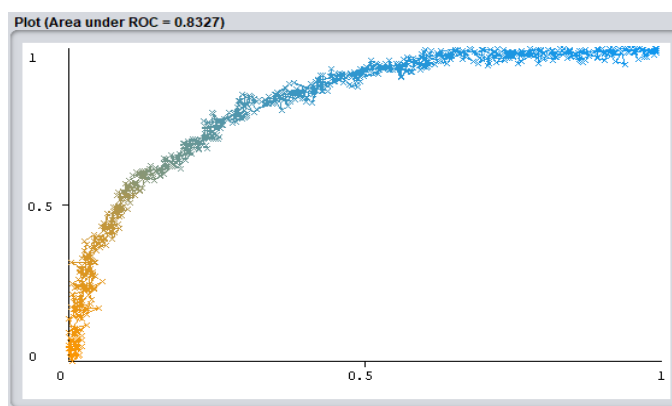
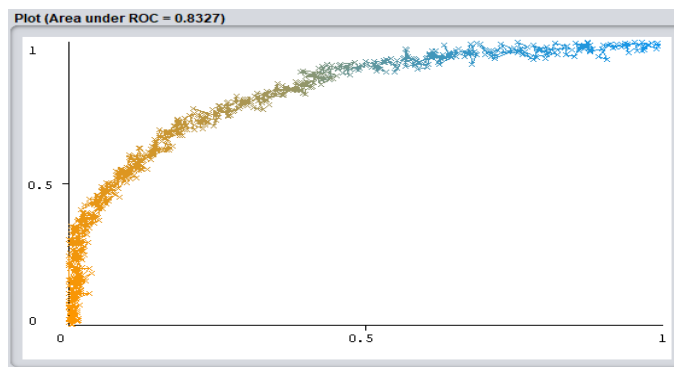


Figure 2. Multiple ROC curve based on diabetes.

Classifier	Area under ROC
KNN	0.6501
SVM	0.7175
J48	0.7512
RF	0.8179
NB	0.8184



Logistic	0.8316
Meta Vote	0.8327

Figure 3. Multiple ROC curve based on Non-diabetes.

In the individual analysis, we represent the individual ROC curve based on diabetes and non-diabetes. But in multiple comparisons, we use six machine learning algorithms. Since the ROC curve of Random Forest 0.818, Naïve Bayes 0.818, Simple logistic 0.832, Meta vote classifier are almost similar. For understanding the above graph clearly, we use only six machine learning algorithms for both diabetes and non-diabetes. And its corresponding ROC area is Support Vector Machine 0.751, KNN 0.650 for both diabetes and non-diabetes. Finally, we see that the cross sign of Meta Vote classifier represents the best ROC curve for both diabetes and non-diabetes.

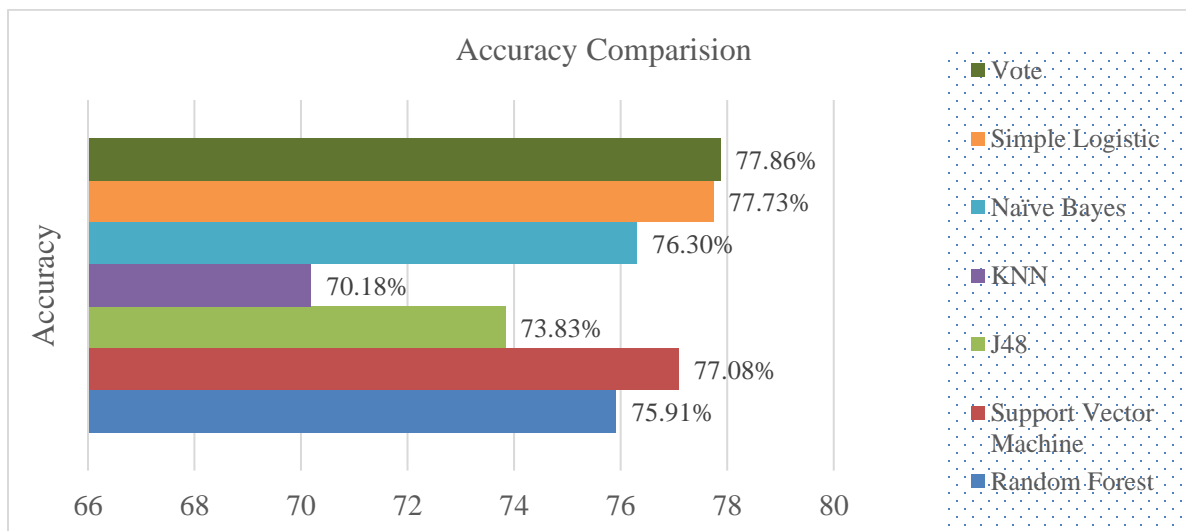


Figure 4. Different Machine Learning algorithm’s accuracy based on diabetes datasets.

The best measure of performance by different Machine Learning algorithms, we can see that Meta Vote classifier which is the combination of Random Forest, Support Vector Machine, J48, Naïve Bayes, Bayes Nets, Simple Logistic represent the best accuracy of 77.8646% and others represent Simple logistic 77.7344%, Naïve Bayes 76.3021%, KNN 70.1832%, J48 73.8281%, Support vector machine 77.0833% and Random forest represent 75.99115% accuracy for correctly classification of diabetes datasets.

5. Conclusion

Diabetes is becoming an epidemic in Bangladesh. Machine learning classifier algorithm is essential for predicting diabetes with better percentages of accuracy. The combination of six base-level classifiers with Meta vote classifier according to the average probability; that are Random Forest, Support Vector Machine, Naïve Bayes, Bayes Nets, J48 and Simple Logistic machine learning algorithm provides higher accuracy for both diabetes and non-diabetes. Since we cannot destruction diabetes forever but we can control diabetes by diet, physical activity and medicine with appropriate treatment. People should increase consciousness for diabetes. For increasing consciousness Government and Non-government organizations should arrange a program that will increase consciousness about diet, physical activity and treatment of diabetes.

Conflict of Interest

No conflict of interest.

REFERENCES

- [1] National institute diabetes and digestive and kidney diseases provides information about Diabetes. <https://www.niddk.nih.gov/healthinformation/diabetes/overview/what-is-diabetes>
- [2] World Health Organization, World Health Organization provides site note about High Blood Glucose, 2020. <https://www.who.int/news-room/fact-sheets/detail/diabetes>
- [3] S. Akter, M. M. Rahman, S. K. Abe, P. Sultana, Prevalence of diabetes and prediabetes and their risk factors among Bangladeshi adults: a nationwide survey, Bull. World Health Organ. 92 (2014) 204-213.
- [4] G. T. Ahmed, The threat of undiagnosed diabetes in Bangladesh, The Daily Star, Dhaka, Nov. 25, 2018. <https://www.thedailystar.net/health/diabetes-treatment-in-bangladesh-the-threat-undiagnosed-1664635>

- [5] World Health Organization, Bangladesh Bureau of Statistics (BBS) estimation on Diabetes, 2016. <https://www.who.int/diabetes/country-profiles/en>
- [6] P. Palma, A worrying picture of diabetes in Bangladesh, The Daily Star, Dhaka, Nov. 14, 2018. <https://www.thedailystar.net/supplements/world-diabetes-day-2018/news/worrying-picture-diabetes-bangladesh-1659979>
- [7] S.M. Abrar Aowsaf, Diabetes management service launched in Bangladesh, The Dhaka Tribune, Dhaka, Sep. 25, 2018. <https://www.dhakatribune.com/bangladesh/dhaka/2018/09/25/diabetes-management-service-launched-in-bangladesh>
- [8] I. Mahbub, The State of Diabetes in Bangladesh, Future Startup, Jul. 27, 2016. <https://futurestartup.com/2016/07/27/the-state-of-diabetes-in-bangladesh>
- [9] N. I. Hasib, Children getting type 2 diabetes 'alarmingly' in Bangladesh, bdnews24.com, Dhaka, Apr. 06, 2016. <https://bdnews24.com/health/2016/04/06/children-getting-type-2-diabetes-alarmingly-in-bangladesh>
- [10] IndexMundi, Diabetes prevalence refers to the percentage of people ages 20-79 who have type 1 or type 2 diabetes, 2019. <https://www.indexmundi.com/facts/indicators/SH.STA.DIAB.ZS/rankings/asia>
- [11] IndexMundi, Diabetes prevalence (% of population ages 20 to 79) - Country Ranking, 2019. <https://www.indexmundi.com/facts/indicators/SH.STA.DIAB.ZS/rankings>
- [12] M. Alehegn and R. Joshi, Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach, *Int. Res. J. Eng. Technol.* 4 (2017) 426–436.
- [13] B. Naqvi, A. Ali, M. A. Hashmi, and M. Atif, Prediction Techniques for Diagnosis of Diabetic Disease: A Comparative Study, *Int. J. Comput. Sci. Netw. Secur.* 18 (2018) 118–124.
- [14] K. P. Ahmed, Analysis of data mining tools for disease prediction, *J. Pharm. Sci. Res.* 9 (2017) 1886–1888.
- [15] Y. Li, H. Li, and H. Yao, Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach in New Urban Area of Urumqi, Xinjiang, China, 2016–2017, *Comput. Math. Methods Med.*, vol. 2018 (2018) 7207151.
- [16] M. Alehegn, R. Joshi, and P. Mulay, Analysis and prediction of diabetes mellitus using machine learning algorithm, *Int. J. Pure Appl. Math.* 118 (2018) 871–878.
- [17] D. J. G. R. S. T. Padma Nivethitha, M. Uma Maheswari, A Survey on Classification and Prediction Techniques in Data Mining for Diabetes Mellitus, *Int. J. Trend Sci. Res. Dev.* 2 (2018) 496–504. <https://doi.org/10.31142/ijtsrd15878>.
- [18] R. Suryakirani and R. Porkodi, Comparative Study and Analysis of Classification Algorithms In Data Mining Using Diabetic Dataset. *Int. J. Sci. Res. Sci. Technol.* 4 (2018) 299–304.
- [19] F. Mercaldo, V. Nardone, and A. Santone, Diabetes mellitus affected patients classification and diagnosis through machine learning techniques, *Procedia Comput. Sci.* 112 (2017) 2519–2528.
- [20] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, Comparison of three data mining models for

- predicting diabetes or prediabetes by risk factors, *Kaohsiung J. Med. Sci.* 29 (2013) 93–99.
- [21] A. kumar Dewangan and P. Agrawal, Classification of diabetes mellitus using machine learning techniques, *Int. J. Eng. Appl. Sci.* 2 (2015) 145–148.
- [22] K. Saravananathan and T. Velmurugan, Analyzing diabetic data using classification algorithms in data mining, *Indian J. Sci. Technol.* 9 (2016) 1–6.
- [23] M. Mounika, S. D. Suganya, B. Vijayashanthi, and S. K. Anand, Predictive analysis of diabetic treatment using classification algorithm, *Int. J. Comp. Sci. Inf. Technol.* 6 (2015) 2502–2505.
- [24] E. G. Yıldırım, A. Karahoca, and T. Uçar, Dosage planning for diabetes patients using data mining methods, *Procedia Comput. Sci.* 3 (2011) 1374–1380.
- [25] N. Chandgude and S. Pawar, A survey on diagnosis of diabetes using various classification algorithm, *Int. J. Recent Innov. Trends Comput. Commun.* 3 (2015) 6706–6710.
- [26] M. Reyaz and G. Dhawan, Various Data Mining Techniques echniques for Diabetes Prognosis rognosis : A Review, *Int. J. Trend Sci. Res. Dev.* 2 (2018) 2–7.
- [27] G. Kaur and A. Chhabra, Improved J48 classification algorithm for the prediction of diabetes, *Int. J. Comput. Appl.* 98 (2014) 13–17.
- [28] S. Sa'di, A. Maleki, R. Hashemi, Z. Panbechi, and K. Chalabi, Comparison of data mining algorithms in the diagnosis of type II diabetes, *Int. J. Comput. Sci. Appl.* 5 (2015) 1–12.
- [29] V. A. Kumari and R. Chitra, Classification of diabetes disease using support vector machine, *Int. J. Eng. Res. Appl.* 3 (2013) 1797–1801.
- [30] T. R. Prajwala, A comparative study on decision tree and random forest using R tool, *Int. J. Adv. Res. Comput. Commun. Eng.* 4 (2015) 196–199.
- [31] J. P. Kandhasamy and S. Balamurali, Performance analysis of classifier models to predict diabetes mellitus *Procedia Comput. Sci.* 47 (2015) 45–51.
- [32] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, Performance analysis of data mining classification techniques to predict diabetes, *Procedia Comput. Sci.* 82 (2016) 115–121.
- [33] Weka, The workbench for machine learning, <https://www.cs.waikato.ac.nz/ml/weka>
- [34] Md. Hossain, Md. Asadullah, A. Rahaman, Md. Miah, M. Hasan, T. Paul, M. Hossain, Prediction on Domestic Violence in Bangladesh during the COVID-19 Outbreak Using Machine Learning Methods, *Appl. Syst. Innov.* 4 (2021) 77. <https://doi.org/10.3390/asi4040077>.