Improved Prediction of Diabetes Mellitus using Machine Learning Based Approach

Madhumita Pal
Dept. *of Electrical Engg.*
Govt.college of Engg.
Keonjhar-758002, Odisha, India
madhumitapal@gcekj.ac.in

Smita Parija
Dept. of Electronics and Telecommunication
Cvaman Global Univesity
Bhubaneswar,Odisha,India
smita.parija@gmail.com

Ganapati Panda
Dept. of Electronics and Telecommunication
Cvraman Global Univesity
Bhubaneswar,Odisha,India
ganapati.panda@gmail.com

Abstract— The diabetes is one of the most commonly occurring chronic diseases in human being. Statistical models are available for prediction of diabetes but these provide poor performance. This article proposed machine learning based model for prediction of diabetes disease. Three supervised machine learning algorithms namely K-NN, Linear SVM and Random Forest have been chosen for diabetes prediction for early diagnosis. The area under the curve and accuracy of each of these models have been obtained using PIMA Indian Diabetes dataset from UCI repository. The comparative results demonstrate that among these three algorithms random forest is the best model in terms of accuracy of 78.57 and AUC of 95.08 for diabetes risk prediction. The contribution of this article will help the healthcare professionals for the early prediction of the disease and taking appropriate treatment. The proposed approach can be applied for detection of other diseases.

Keywords—Machine leaning, Diagnosis of Diabetes disease, Random Forest, Linear SVM, K-NN.

I. INTRODUCTION

With the growth of population providing proper treatment to patient and of remote area is a challenging task for the healthcare professionals. Hospitals generates out 8 huge amount of patient data and analysing, processing of these data to arrive at a conclusion is a challenging task. Discovering useful diagnosis information from the patient data can help the medical practitioners to diagnose the disease at an early stage. Diabetes is a disease in which blood glucose or blood sugar levels are too high. Glucose comes from the food human eat. It is one of the most dangerous chronic diseases which occurs when the pancreas is unable to produce sufficient insulin. Insulin is a hormone that helps the glucose to get into human cells to produce energy. Over time the diabetes damages
the human organ. The fatality rate due to this chronic disease is less in high income countries than the lower and middle income countries[1]. According to WHO survey report the number of individuals affected with diabetes have increased from 108 million in 1980 to 422 million in 2014[2]. Major side effects of diabetes are Hyperglycaemia blindness, kidney damage, congestive heart failure and even the need to remove a limb. During 2000 and 2016, 5% increase in unseasonable fatality from diabetes have been witnessed. In 2019, nearly 1.5 million death have been caused by diabetes and nearly 2.2 million deaths have been caused due to high blood glucose in 2012.For prevention of diabetes a healthy diet, daily physical exercises, continuing a normal body weight and quieting of tobacco are needed. Diabetes is curable and its consequences avoided or delayed with diet, physical activity, medication and regular screening and treatment for complications.

A. Types of Diabetes

i)Type 2 diabetes

Type 2 diabetes is the more common type diabetes which occurs usually in adults if the body does not make or use insulin well. Without enough insulin level the glucose stays in human blood and over time having too much glucose in blood can cause serious damages to the most of the organs of human body. It mainly occurs due to heavy body weight and lack of physical exercises. This type of diabetes cannot be diagnosed at early stages [4].

ii)Type 1 diabetes

In Type 1 diabetes the body cannot produce insulin. It occurs both in child and in the juvenile. Loss of vision, loss of weight, excess creation of urine, hugeness are the common symptoms associated with type-1 diabetes.

iii)Gestational diabetes

This type of diabetes commonly occurs in women during pregnancy and leads to complications during delivery time of women. In this type of diabetes glucose values remains above normal but below the diagnostics level. Gestational diabetes increased the probability of risk that the pregnant women and their children may have type 2 diabetes in future. Prenatal screening method is used for detection of Gestational diabetes [4].

iv)Impaired glucose tolerance and impaired fasting glycaemia

These are the progressive stage of people towards type 2 diabetes. These are observed between normal and diabetes persons.

Effects of diabetes on health

Over time, the diabetes can damage the heart, blood vessels, eyes, kidneys, and nerves.

- It increases the risk of heart disease two to three times in adults.
- It damages the human nerve system.
- It causes retinopathy which makes people blind forever [2].

2021 2nd International Conference on Range Technology (ICORT) | 978-1-6654-4956-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICORT52730.2021.9581774

- It is the major cause of kidney failure *[3]*.

1.2 Prevention

To prevent the occurrence of type 2 diabetes people should

i) maintain a proper body weight

ii) physical exercises have been needed to fit the body at least 30 minutes within a day.

iii)By eating healthy food and avoiding fat food and sugar it can be prevented.

iv) By avoiding tobacco and smoking

1.3 Diagnosis and treatment

Diagnosis of diabetes can be carried out through blood sugar testing. Treatment of diabetes includes proper diet and physical exercises. Type 1 diabetes can be controlled by controlling the blood glucose level which is by injecting insulin. But treatment of diabetes required oral medication and at times insulin may be required. Thus, by controlling blood pressure and taking care of body weight properly type-2 diabetes can be controlled.

II. Literature Survey

The main objective of this paper is to develop a prediction model for early prediction of diabetes using three supervised machine learning algorithms and comparing their prediction accuracies.

The organization of the paper is as follows. Section II represents the literature survey part of the article. Section III explains the dataset and feature to be used for detection task. Section IV presents the schematic model. Section V explains the machine learning algorithms used for prediction of the disease. Section VI deals with the analysis of results part of the model result. Section VII provides the conclusion part of the article.

Hasan [8] et al have proposed a model for early prediction of diabetes using machine learning models such as k-nearest neighbour, decision Trees, random forest, adaboost, Naïve Bayes, XG Boost and multilayer perceptron. They have used different data pre-processing methods such as removal of outlier, checking null values, selection of attributes and used K-fold cross validation for testing the model. They have obtained the best performance of the model using ensemble classifier having sensitivity, specificity, rate of omission, diagnosis odd ratio and AUC values of 0.78,0.93,0.09,66.23 and 0.95 respectively. In [9] DT, SVM, and NB are used for prediction of diabetes. They have obtained maximum AUC of 0.819% with Naïve Bayes algorithm's.An recent work [10] proposed a diabetes risk prediction models by combining support vector machine and genetic algorithm for providing a decision support to medical practitioners to take decision on discharging a patient from hospital with diabetes. Their model reduces the risk of readmission of patient into the hospital and also prevent the wastage of medical resources and expenses. They have conducted the experiment from 8756 medical samples with 50 features and have obtained accuracy of 81.02%, sensitivity of 82.89%, specificity of 79.23% from their experiment using support vector machine. Arun [11] et al have used four ML classifier such as DT, ANN, LR, and NB to classify the risk of diabetes mellitus.They have found that random forest yields the best result in terms of prediction accuracy. In[12]the authors have a model for prediction of heart disease using random forest algorithm with an accuracy value of 86.9%. Different machine learning models for dimensionality reduction are suggested in [13]. The authors used ML techniques such as (LDA) [14], (QDA) [15], Naive Bayes (NB) [16], Gaussian Process Classification [17], (SVM) [18], (ANN) [19], AdaBoost (AB) [20], Logistics regression (LR) [21], (DT) [22] and k-fold cross-validation scheme for classification purpose. They obtained a maximum AUC of 0.93 by boosting the Machine learning model.

III. Details of Data Set Used

For the experimental simulation study PIMA Indian dataset from UCI repository site is used.It consists of 769 data samples with 9 features which is explained in table1.

Table 1. Description of features used in the dataset

Sl no.	FEATURES	DESCRIPTION
1.	Pregnant	how many times pregnant
2.	Glucose	Glucose concentration of plasma within 2 hours
3.	Pressure diastolic BP	Blood pressure
4.	Skin thickness	Thickness of skin fold
5.	Insulin	Serum insulin
6.	BMI	Body mass index
7.	Age	Age of patient in years
8.	Diabetes pedigree	Pedigree function of diabetes

The preprocessing of data is carried out which includes checking null values, cleaning data, data filtrations and removal of outliers. *The proposed algorithms are implemented in Jupyter environment using python open-source software. Different packages such as pandas, numpy, sklearn, matplotlib are loaded for execution of these algorithms. The window 10, i5 operating environment is used for implementation of the proposed algorithms.*

IV.PROPOSED SYSTEM

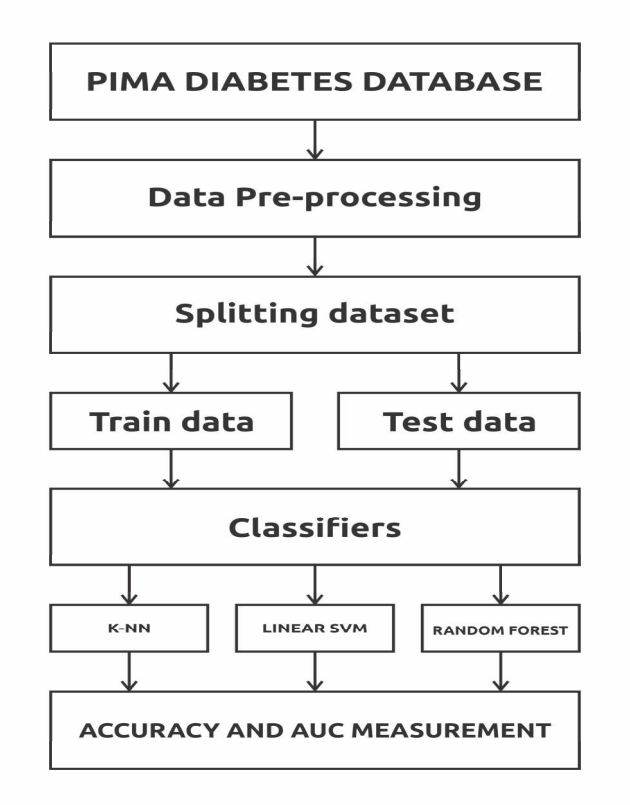

Figure 1.Generalised block diagram of proposed prediction model

The prediction model of Fig.1. is implemented in the following steps.

i). The PIMA diabetes dataset containing 769 data samples of diabetes patients.

ii). Features are selected which are needed for diabetes prediction.

iii). The dataset is examined to find null values and outliers are removed

iv).The 80% feature sets have been used for training and 20% for testing samples.

v). Different classification techniques are simulated using pre-processed data for diabetes prediction

vi). Performance matrices of these models are obtained in terms of accuracy and AUC.

V.PROPOSED MACHINE LEANING ALGORITHMS

The algorithms are explained in concise manner

A.K-NN

This algorithm is one of the simplest supervised machine learning algorithms used for classification. It classifies the new objects based on the distance metric. The class belongingness of the new object depends upon its neighbors maximum voting condition. Euclidean distance metric used for finding the neighbor close to the new objects [4].

B. Support Vector Machine

This algorithm has been introduced by Corinna Cortes and Vladimir Vapnik used for solving both classification and regression problems [5].The support vector machine creates a hyperplane in a high dimensional space and classifies the objects into different classes with maximum margin as shown in Fig 2. The SVM operates both in continuous and discrete variables [4]. It produces a model which plots the training vectors in multi-dimensional space and classifies each training vector by its class [8].

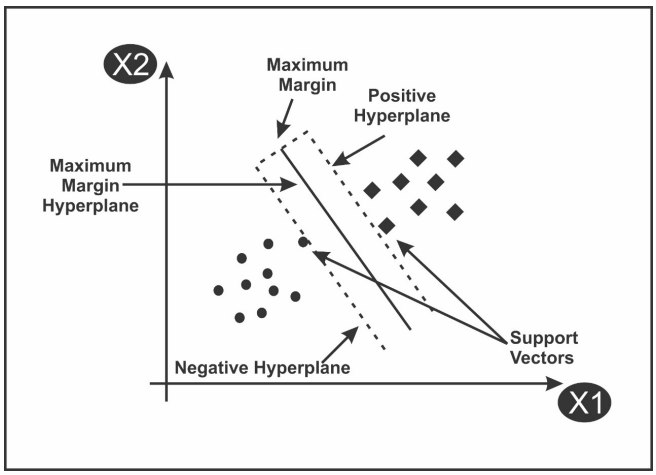

Figure 2. support vector machine classifier

C. Random Forest

It is one of the powerful supervised machine learning models used for classification. This classifiers ensemble the results of many models to form a strong predictive model. This model is less prone to overfitting problem [23]. Random forest uses bagging technique to increase the no. of decision trees in the forest.

1.If the data samples contain M training vectors, then the training samples are chosen randomly from the original data samples. Selected training samples are used for the growth of the forest.

2.For an P number of attributes the best one is selected for splitting the nodes of decision trees. The value remains constant during the splitting process.

VI. Simulation Based Experiment and Analysis of Results

KNN, SVM, and RF ML models are simulated using PIMA Indian diabetes dataset. The dataset consists of 768 samples with 8 attributes.80% of the data samples have been used for training purposes and remaining 20% of the data samples were used for testing the model. Data visualization of glucose level with target is shown in Fig.3. It shows that as the glucose level of person increases the chances of diabetes also increases. Performance of each of these models is obtained in terms of accuracy and the AUC. Classification results confusion matrix, accuracy and AUC score of these models are presented in Tables 2 and 3.The correlation between each variable is explained in Fig .4.

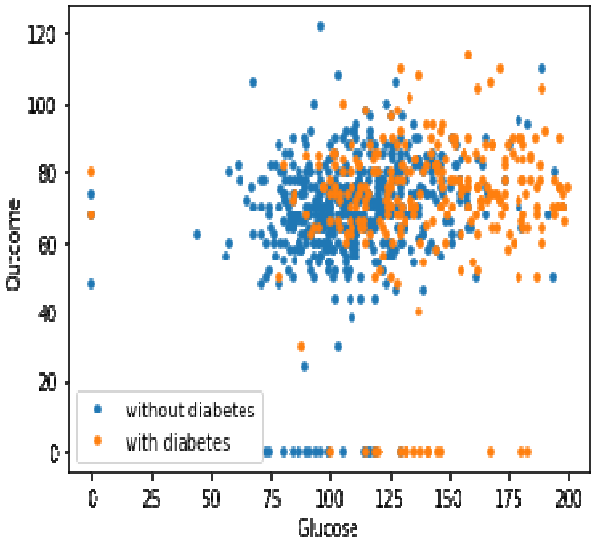

Figure 3.Plot of Data visualisation

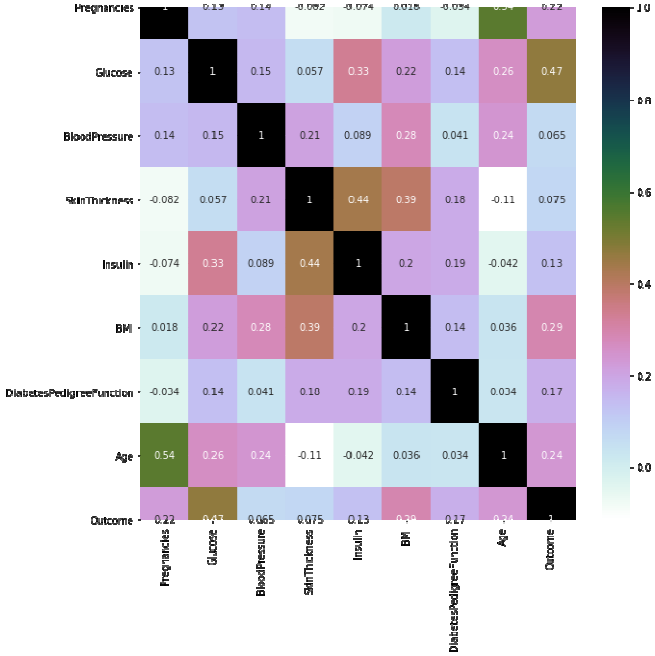

Figure 4.Correlation matix between variables

Table 2.Comparison of confusion matrices of different models

Machine learning algorithms	TP (true positive)	TN (true negative)	FP (false positive)	FN (false negative)
K-NN	21	98	10	25
Linear-SVM	31	89	10	24
Random forest	22	94	6	25

Table 3. Comparison of accuracy and AUC obtained from models

Machine learning Algorithms	Accuracy score	AUC score
K-NN	77.27	68.19
Linear-SVM	77.92	73.13
Random forest	78.57	95.08

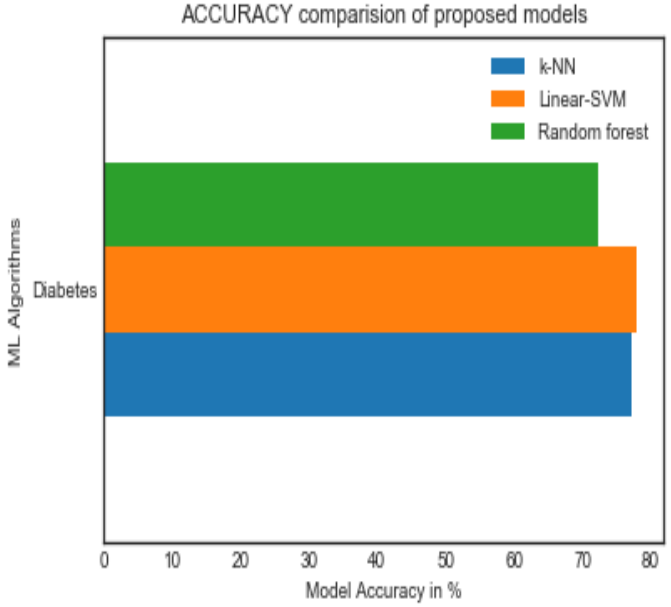

Figure 5. Comparison of accuracy values of proposed models

Fig.5. presents the comparison of accuracy values obtained from the three models. The observation of the plots demonstrates that random forest offers the highest prediction accuracy of 78% followed by SVM and K-NN with an accuracy of 77.92% and 77.27% Rrespectively.

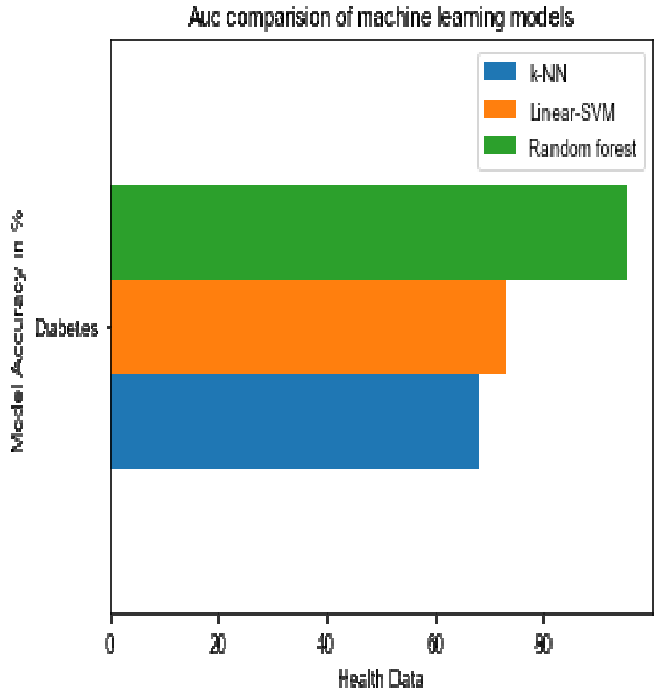

Figure 6. Comparison of AUC values of proposed models

The plots of Fig. 6. clearly exhibits that by using random forest highest AUC of 95% is achieved followed by Linear SVM and K-NN with an AUC values of 73.13% and 68.19% respectively.

Table 4. Performance comparisons of proposed work with the existing work

publication	Algorithms	Obtained accuracy/auc
Mani et al., 2012[23]	Random Forest (RF)	AUC=0.877
Sneha et.al.,2019[24]	K-NN	Accuracy=63.04
Sneha et. al,2019[24]	Random forest	Accuracy=75.39
Sisodia et al.,2018[9]	Support vector machine	Accuracy=65.10, ROC=0.500
Proposed work	K-NN	Accuracy=77.27%, AUC=68.19%
	Linear SVM	Accuracy=77.92%, AUC=73.13
	Random Forest	Accuracy=78.57, AUC=95.08%

Performance comparisons of different algorithms with accuracy and AUC values are compared with the proposed work has been shown in table 4.The above table shows that the performance of our model is better as compared to other existing work in terms of accuracy and AUC.

VII. Conclusion

This article develops and finds the performance measures of three machine learning models for prediction of diabetes . From the experimental study the highest prediction accuracy of 78.57% is achieved by the random forest model followed by linear SVM with an accuracy of 77.92% and K-NN with accuracy of 77.27%. The maximum AUC of 95.08% is obtained from random forest model followed by linear SVM and K-NN with an AUC value of 73.13% ,68.19% respectively. In future the proposed models can be employed for prediction of other diseases such as cardiovascular disease, breast cancer, dermatology to achieve higher prediction accuracy. By integrating IoT technology with machine learning doctors can monitor the glucose level of remote area diabetes patient. Future extension of the current work can be made by applying these models to other data sets to the potentiality and robustness.

References

[1] Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Emerging Risk Factors Collaboration.Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, Di Angelantonio et al. Lancet. 2010; 26;375:2215-2222.

[2] Causes of vision loss worldwide, 1990-2010: a systematic analysis. Bourne RR, Stevens GA, White RA, Smith JL, Flaxman SR, Price H et al. Lancet Global Health 2013;1:e339-e349

[3] 2014 USRDS annual data report: Epidemiology of kidney disease in the United States.United States Renal Data System. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2014:188–210.

[4] J. Pradeep Kandhasamy*, S. Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus", ScienceDirect, Procedia Computer Science 47 (2015) 45 – 51

[5] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[6] Colin Campbell and Yiming Ying, Learning with Support Vector Machines, 2011, Morgan and Claypool. Available: http://www. morganclaypool.com/doi/abs/10.2200/S00324ED1V01Y201102 AIM010?journalCode=aim

[7] H.Barakat, Andrew P.Bradley and Mohammed Nabil H.Barakat (2009) "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus", IEEE Transactions on Information Technology in Bio Medicine, Volume 14, Issue 4, pp 1-7, 2009. Available: http://ieeexplore. ieee.org /xpls/ abs_ all.jsp?arnumber=5378519 Digital Object Identifier: 10.1109 /TITB. 2009.2039485.

[8] Hasan.k, Alam.A ,Dola.D, Hossain.E; Hasan.M, Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers,IEEETrans.volume 8, pp.2169-3536,April 2020.

[9] D. Sisodia and D. S. Sisodia, ''Prediction of diabetes using classification algorithms,'' Procedia Comput. Sci., vol. 132, pp. 1578–1585, Jan. 2018.

[10] Cui.S , Wang .D, Wanga.Y , Wen Yuc.P , Jin.Y "An improved support vector machine-based diabetic readmission prediction" Computer Methods and Programs in Biomedicine,VOLUME-166,2018,Pp 123-135

[11] N. Nai-arun and R. Moungmai, ''Comparison of classifiers for the risk of diabetes prediction,'' Procedia Comput. Sci., vol. 69, pp. 132–142, Dec. 2015.

[12] M. Pal, S.Parija, Prediction of Heart Diseases using Random Forest,Journal of Physics: Conference Series, 1817 (2021) 012009, doi:10.1088/1742-6596/1817/1/012009M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[13] M. Maniruzzaman, M. J. Rahman, M. Al-MehediHasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, ''Accurate diabetes risk

stratification using machine learning: Role of missing value and outliers,'' J. Med. Syst., vol. 42, no. 5, p. 92, May 2018.

[14] G. J. McLachlan, ''Discriminant analysis and statistical pattern recognition,'' J. Roy. Stat. Soc., Ser. A, Statist. Soc., vol. 168, no. 3, pp. 635–636, Jun. 2005.

[15] T. M. Cover, ''Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition,'' IEEE Trans. Electron. Comput., vols. EC–14, no. 3, pp. 326–334, Jun. 1965.

[16] G. I. Webb, J. R. Boughton, and Z. Wang, ''Not so naive bayes: Aggregating one-dependence estimators,'' Mach. Learn., vol. 58, no. 1, pp. 5–24, Jan. 2005.

[17] S. Brahim-Belhouari and A. Bermak, ''Gaussian process for nonstationary time series prediction,'' Comput. Statist. Data Anal., vol. 47, no. 4, pp. 705–712, Nov. 2004. [12] C. Cortes and V. Vapnik, ''Support-vector networks,'' Mach. Learn., vol. 20, pp. 237–297, Sep. 1995.

[18] A. Reinhardt and T. Hubbard, ''Using neural networks for prediction of the subcellular location of proteins,'' Nucleic Acids Res., vol. 26, no. 9, pp. 2230–2236, May 1998.

[19] B. Kégl, ''The return of AdaBoost.MH: Multi-class Hamming trees,'' 2013, arXiv:1312.6086. [Online]. Available: http://arxiv.org/abs/1312.6086

[20] B. P. Tabaei and W. H. Herman, ''A multivariate logistic regression equation to screen for diabetes: Development and validation,'' Diabetes Care, vol. 25, no. 11, pp. 1999–2003, Nov. 2002. [21] I. Jenhani, N. B. Amor, and Z. Elouedi, ''Decision trees as possibilistic classifiers,'' Int. J. Approx. Reasoning, vol. 48, no. 3, pp. 784–807, Aug. 2008.

[21] L. Breiman, ''Random forests,'' Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001

[22] G.Dinesh, K.Arumugaraj, D.Santosh, V.Mareeswari" Prediction of Cardiovascular Disease Using Machine Learning Algorithms" Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India

[23] S. Mani ,Y. Chen ,T. Elasy ,W. Clayton , J.Denny "Type 2 diabetes risk forecasting from EMR data using machine learning". AMIA Annu Symp Proc. 2012; 2012:606-15. Epub 2012 Nov 3.

[24] N. Sneha ,T. Gangil ,"Analysis of diabetes mellitus for early prediction using optimal features selection" *J Big Data (2019) 6:13* https://doi.org/10.1186/s40537-019-0175-6