GIET UNIVERSITY JOURNAL, 2022, VOLUME 1, ISSUE 1.
Manuscript received on 2nd July 2022: Revised on 16th July, 2022; Accepted on 26th July, 2022

Interactive Machine Learning Classification Techniques for Diabetes Mellitus Prediction

[1]Sachikanta Dash, [2]Sasmita Padhy

Department of CSE, [1]GIET University,Gunupur, [2]VIT Bhopal University

***Abstract*—Diabetes is a chronic condition with the potential to destabilize the global health-care system. Diabetes affects 382 million people globally, according to the International Diabetes Federation. By 2035, this number will have risen to 592 million. Diabetes is a condition that causes blood glucose levels to rise. Frequent urine, increased thirst, and increased appetite are all signs of high blood glucose. Blindness, renal failure, amputations, heart failure, and stroke are just a few of the diabetic complications. Our bodies convert food into sugars, or glucose, when we eat it.Machine learning is a relatively young branch of data science that studies how computers learn from their past experiences. The goal of this project is to use a mix of machine learning approaches to build a system that can identify diabetes in a patient earlier and more accurately. Support Vector Machine, Logistic Regression, Random Forest, and K-Nearest Neighbor are four supervised machine learning algorithms used in this work to predict diabetes.The accuracy of the model is calculated using each algorithm. After that, the model with the highest accuracy for predicting diabetes is chosen. A comparison analysis is proposed in this work for properly predicting diabetes mellitus. This study also attempts to provide a more effective method for detecting diabetes illness.**

***Keywords*—K Nearest Neighbor, Logistic Regression, Support Vector Machine, Random Forest, ROC**

I. INTRODUCTION

THISarticle requires certain precise actions related to diabetes control and prevention. Previously, statistics showed that one out of every 10 people in the United States had diabetes. Regardless, it is expected that by 2045, it will be able to assist one out of every three people. This is a real problem that has to be addressed. When the blood glucose concentration rises to dangerously high levels, diabetes becomes a chronic condition. This is a major cause of various problems and illnesses, such as renal disease and heart disease. Diabetic predisposition is also caused by a variety of bad food habits and a lack of proper bodily routines. The World Health Organization (WHO) has said that the total number of people living with diabetes has skyrocketed in recent years. Managing multiple diabetic datasets is required to enhance the present rate of diabetes patients and reduce it to an absolutely inconsequential level by focusing on reducing it on a large scale. Certain approval techniques are also incorporated to operate with the diabetes forecast project's pure precision.

Diabetes is a rapidly spreading disease that affects individuals of all ages, including children. To understand diabetes and how it develops, we must first understand what happens in the body without diabetes. Carbohydrate meals are our body's primary source of energy. Bread, cereal, pasta, rice, fruit, dairy products, and vegetables are all carbohydrate foods (especially starchy vegetables). When we ingest these nutrients, it is then converted to glucose by our body. The glucose then travel through the bloodstream in our body. Few glucose particle is also transported to our brain, that helps us and giving us the ability for operate and thinking properly. The rest of the glucose particle then sent to the body's cells for energy. The remaining are sent to liver for latter use of energy. Insulin is necessary for the body to utilize glucose for energy. Insulin is a hormone that is generated in the pancreas by beta cells. Insulin works in a similar way as a key in a lock. Insulin attaches to cell doors and opens them, allowing glucose to flow from the circulation into the cell through the cell door. When the pancreas is unable to create enough insulin or the body is unable to use the insulin it does make (insulin resistance), glucose builds up in the bloodstream, causing hyperglycemia and diabetes. The presence of excessive quantities of sugar (glucose) in the blood and urine is a symptom of diabetes mellitus.

A. Type of Diabetes with Symptoms

The diabetes mellitus can be categories to following three types [1]:

- Type-1 diabetes is defined by the pancreas producing less insulin than the body requires, a condition known as "insulin-subordinate diabetes mellitus" (ISDM). Type-1 diabetics require supplemental insulin to compensate for the pancreas' decreased insulin production.
- Type-2 diabetes is defined as insulin resistive body, which occurs when the body’s cells react to the insulin differently than they would ordinarily. "Adult starting diabetes" or "non-insulin subordinate diabetes mellitus" (NISDM) are other terms for this condition. This kind of diabetes is more common in those with a high BMI or who have a sedentary lifestyle.
- During the time of pregnancy, the third type of diabetes called Gestational diabetes may develop.

A typical human's sugar levels may vary from range 70 to 99 milligrams per deciliter. A person is classified as diabetes when her or his fasting glucose level reached to 126 mg/dL. In medical point of view, someone having a glucose level of 100 to 125 mg/dL may be considered as pre-diabetic [2]. In such an individual, type 2 diabetes is more prone to develop. GDM (gestational diabetes mellitus) is a kind of diabetes that develops during pregnancy that is not clearly evident diabetes during 2nd and 3rd trimester of pregnancy. Diabetic may be caused by other factors, such as monogenic diabetes syndromes, exocrine pancreas diseases.

Symptoms:The symptoms for diabetes may vary depending on blood glucose level. Some people, particularly those with type-2 diabetes or prediabetes, may not show any signs at all. Symptoms of type-1 diabetes appear more quickly and are more severe.Some of the signs and symptoms of type 1 and type 2 diabetes are as follows:

- Urination on a regular basis
- An increase in thirst
- Tired/Sleepiness
- Loss of weight
- Distorted eyesight
- Emotional ups and downs
- Perplexity and inability to concentrate
- Infections are common

Diabetes has a number of causes*:* Genetic factors account for the majority of diabetes cases. At least two faulty genes on chromosome, the chromosome that controls the body's antigen response, are to blame. The development of type 1 and type 2 diabetes has been related to viral infection. Rubella, Coxsackievirus, mumps, hepatitis B virus, and CMV infection have all been related to an increased risk of diabetes.

II. Literature Survey

This section looks at a few works that are linked in some way. Numerous scientific research have utilised the Pima Indians Dataset for Diabetes (PIDD) to predict diabetes. Weka and machine learning approaches were used in [3]. Data mining, Machine learning, neural network, and hybrid techniques are among the methodologies used by researchers.

Swapna et al. in [4] used electrocardiogram (ECG) data to detect diabetes using deep learning algorithms. They retrieved features using a convolution neural network (CNN) then support vector machine algorithm is used to extract the features. Finally, they determined that the accuracy rate was 95.7 percent. In this healthcare area, a variety of computing approaches were utilized. The application of several machine learning algorithms for predicting diabetes mellitus is the subject of this literature review. We extract information from the provided medical data in order to achieve flawless accuracy. Md. Faisal Faruque et al. [5] used the random forest method to create a predictive analytic model. Jyotismita Chaki et al. [6] utilized 10 fold cross validation as an assessment technique for three distinct algorithms: decision tree, naive bayes, and SVM, with naive bayes outperforming the other algorithms by 75 percent.

To forecast diabetes mellitus in its early stages, Choi, B.Get al.[7] utilized random forest, KNN, nave bayes, SVM, and decision tree. We are now using machine learning algorithms and statistical data in the healthcare area to comprehend the sick data that has been discovered. Because the machine learning area encompasses a wide range of approaches and studies, it's difficult to establish a comparison based on which algorithm is faster at producing prediction results. The algorithm's categorization was not tested using the cross validation approach. Different data mining approaches were utilised to predict and assess diabetic mellitus. We used real-world data sets by gathering information from the supplied datasets since we employed three data mining approaches.

K.VijiyaKumar et al. [8] presented the Random Forest method using machine learning techniques. The suggested model produces the best diabetic prediction results, demonstrating quickly predicting diabetes mellitus.

Predicting diabetes onset: an ensemble supervised learning technique was reported by Nonso Nnamoko et al. [9]. For the ensembles, five commonly used classifiers are utilised, and their outputs are aggregated using a meta-classifier. The findings are reported and compared to other research in the literature that used the same dataset. It is demonstrated that diabetes onset prediction can be done more accurately utilising the proposed technique.

Diabetes Prediction Using Machine Learning Techniques, given by Tejas N. Joshi et al. [10], seeks to predict diabetes using three distinct supervised machine learning methods: SVM, Logistic regression, and ANN. This study provides an excellent method for detecting diabetes illness sooner. Sisodia, Det al. [11] proposed employing data mining to build an Intelligent Diabetic Disease Prediction System that provides analysis of diabetes malady using a database of diabetes patients. In this approach, they suggest using Bayesian and KNN (K-Nearest Neighbor) algorithms to a diabetes patient database and analysing it using multiple diabetes characteristics to forecast diabetes illness.

Muhammad Azeem Sarwar et al. [12] used four different machine learning algorithms indicating which algorithm is best algorithm for diabetes prediction. Researchers are interested in diabetes prediction in order to train a software to determine if a patient is diabetic or not by using a suitable classifier on a dataset. The categorization procedure, according to prior studies, has not much improved. As Diabetes Prediction is an important topic in computers, a system is necessary to tackle the difficulties highlighted based on past research.

We examined actual diagnostic medical data based on numerous risk variables in order to classify machine learning techniques and forecast diabetes mellitus in this study.

III. Various Machine Learning Approaches

To predict diabetes, we employ a variety of classification methods. This is a crucial characteristic that plays a big part in prediction. The methods are as follows:

A. Logistic Regression:

In 1958, statistician DR Cox invented logistic regression, which predates the area of machine learning. It's a type of

supervised machine learning approach used in classification tasks (for predictions based on training data). Logistic Regression employs the same equation as Linear Regression, however the result is a categorical variable in logistic regression, whereas it is a value in other regression models. The independent variables can be used to predict binary outcomes. The Logistic regression model is a form of machine learning classification model that has the binary valueslike 0 or 1, -1 or 1, true or false as dependent variable and the independent variable like interval, ordinal, binominal or ratio-level[13]. The logistic/sigmoid equation function is as follows

$$y = \frac{1}{1+e^{-x}} \quad (1)$$

B. K nearest neighbor:

Both classification and regression issues may be solved using the K-Nearest Neighbor (KNN) technique [14]. However, in the industry, it is more commonly utilised in classification issues. KNN is a straightforward computation that stores all existing examples and ranks new ones based on the votes of its k neighbours. To place the case in the class with the most people among its K nearest neighbours, a distance work is used. The Manhattan, Hamming, Euclidean, and Makowski distances are among the distance capabilities. The first 3 nos. of features are used for indefinite functions, whereas the 4th one is used for absolute variables. If K = 1, the case is essentially assigned to the class of the next-closest neighbour. Selecting K for KNN modelling might be challenging at times. Its main benefit is the ease with which it may be translated and the little amount of time it takes to compute

C. *Support Vector Machine (SVM) Classifier:*

SVM is a supervised machine learning method that excels in pattern identification and is used as a training process for deducing classification and regression rules from data. When the number of characteristics and instances is large, SVM is the most exact method. The SVM algorithm creates a binary classifier. In an SVM model, each data item is represented as a point in an n-dimensional space, with n being the number of features, and each feature as the value of a coordinate in the n-dimensional space. The basic objective of SVM is to use an appropriate hyper plane to categorize data points in a multidimensional space. A hyper plane is considered as a boundary of classification for data points. The hyper plane classifies the data points with the biggest gap between the classes and the hyper plane. In this technique, each data item in n-dimensional space is represented as a point, with the value of each feature matching to the value of a certain coordinate.Because the two closest focuses are the furthest distant from the line in figure 1, the dark line divides the data into two different organized groupings. Our classifier is represented by this line. Based on falling of testing data on both side of the line, the new data be able to categorized into one of two categories.

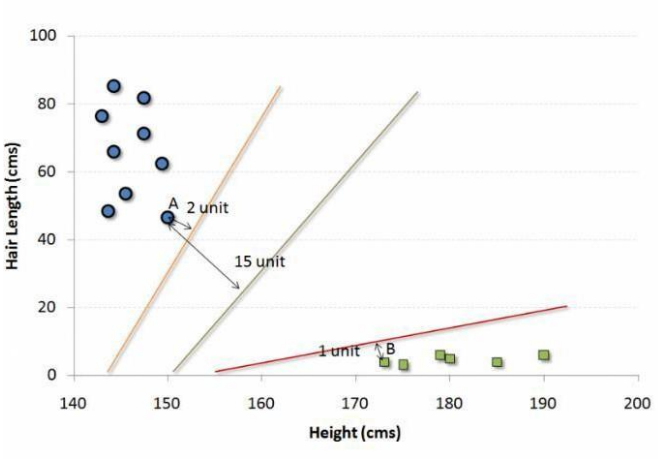

Fig. 1. Support Vector Machine

D. Random Forest (RF):

The RF is made up of several separate decision trees that operate together as an ensemble, as the name indicates. For each tree, the RF predicts a class, and the classes with the majority of votes become our model's prediction report. Random Forest is another frequently used supervised machine learning technique. This method works well for both regression and classification issues, although it excels at the latter. As the name indicates, the Random Forest technique analyses many decision trees before providing an output. As a consequence, it's a decision tree collection. This technique is based on the assumption that if additional trees were present, they would all reach the same conclusion. For classification, it utilizes a voting technique and then selects the class, but for regression, it takes the mean of all decision tree outputs. It works well with large datasets with several dimensions.

IV. METHODOLOGY

This section will cover the various classifiers used in machine learning to predict diabetes. We'll also go through our recommended technique for increasing accuracy. In this article, five alternative techniques were employed. The many methods utilized are listed below. The accuracy measurements of the machine learning models are the output. The model may then be utilized to make predictions.

A. Dataset Description

The diabetes data set is downloaded from Kaggle repository[15]. A diabetes dataset of 2000 cases were used. The diabetes dataset of Pima Indians was used to test the techniques.The goal is to determine whether or not the patient is diabetic based on the measurements.Parameters used in Pima datasets are; 1) Age 2) Glucose 3) BloodPressure 4) BMI 5) Insulin 6) SkinThickness 7) DiabetesPedigreeFunction 8) Pregnancies 9) Outcome. Figure 2 depicts description matrix of Pima Indian dataset and figure 3 depics hit map of the correlation matrix of the same.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Fig.2. Screen shot of description of matrices of Pima dataset

Fig.3. Hit map of correlation matrix for the dataset

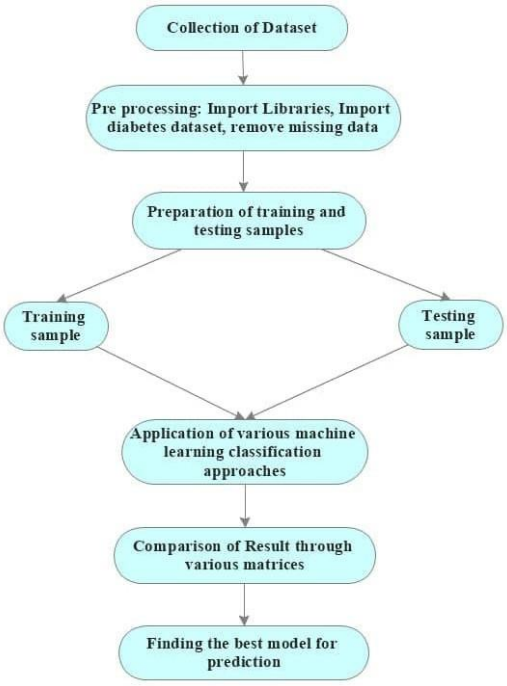

Fig.4. Flow graph for the proposed work

B. Implementation and Design

The study's implementation was done with Google Colab, and the coding was done with the python programming language. The Pima dataset and the gathered dataset were used to forecast availability of diabetes using various machine learning approaches such as SVM , k-nearest neighbour, RF and LR classifications. After collection of dataset, the missing values are checked as in table 1. After then, each classifier's predictions are compared to one another. The procedures to implement the machine learning algorithm are illustrated in Figure 4.

The data set used to predict diabetes is shown in Figure 2. The diabetes parameters serve as the variable which is dependent, whereas the other factors served as independent ones. For the dependent diabetes features only two values are accepted, with a "zero" indicating No diabetes and a "One" signifying availability of diabetes. The whole sample is divided into two groups, with a ratio of 70:30 for the training and testing dataset. All four methods of classification, i.e. were used for prediction. The training data was then used to predict the test set outcomes using SVM , k-nearest neighbour, RF and LR classifications, resulting in the confusion matrix given in Table 2.

The measure provided in equation 2-8 may be computed using the obtained confusion matrices. True Negative (TN), False Negative (FN), True Positive (TP), and False Positive (FP) were the results of these matrices (TP). Because there are more non-diabetic cases than diabetic cases in both datasets, the TN is greater than the TP. As a consequence, all of the techniques provide positive results. The following measurements have been calculated using the following formulae in order to determine the precise accuracy of each method:

TABLE I
FINDING MISSING VALUES IN DATASET

Properties	Missing Values
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

TABLE II
MATRIX OF CONFUSION FOR DIFFERENT CLASSIFICATION METHODS

	Logistic Regression	K Nearest Neighbour	Support Vector Machine	Random Forest
Pima Dataset	[[149 11] [31 40]]	[[141 19] [28 43]]	[[142 8] [46 35]]	[[131 1] [2 97]]

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TP+FP} \quad (4)$$

$$\text{MCC} = \frac{(TP*TN)-(FP+FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

$$\text{ErrorRate} = \frac{FN+FP}{TP+TN+FN+FP} \quad (6)$$

$$\text{F-Measure} = \frac{2*(Precision*Sensitivity)}{Precision+Sensitivity} \quad (7)$$

$$\text{Accuracy} = \frac{TN+TP}{TP+TN+FN+FP} \quad (8)$$

TABLE III
COMPARISON OF STATISTICAL MEASUREMENT FOR VARIOUS CLASSIFICATION TECHNIQUES

	Logistic Regression	K Nearest Neighbour	Support Vector Machine	Random Forest
Accuracy	**0.872**	**0.739**	**0.888**	**0.984**
Error	0.127	0.261	0.112	0.016
Sensitivity	0.923	0.778	0.898	0.987
Specificity	0.764	0.702	0.816	0.916
Precision	0.885	0.816	0.933	0.991
F-Measure	0.903	0.797	0.915	0.989
MCC	0.732	0.503	0.764	0.963
Kappa	0.727	0.516	0.713	0.922
AUC	**0.908**	**0.916**	**0.893**	**1**

Another finding as per table 3 is that the accuracy level of all the techniques is higher on our collected dataset than on the used PIMA dataset, owing to the former'sgreater number of variables relevant to assessing diabetes risk. The Random Forest classifier outperforms all others in terms of accuracy (98.4%), sensitivity, specificity, precision, and F-measure, proving that it is the best technique for our dataset. Furthermore, in the case of random forest, the AUC value is 1, indicating that this model performs exceptionally well in classification. Figure 5 Depicts the clear graph for the ROC curve and AUC the PIMA datasets and figure 6 depicts the Accuracy graph for the dataset. A comparison graph is presented in figure 7 that clearly compare all the available algorithms. Here it indicates that in the cases the RF classifier gives the highest result with value 1.

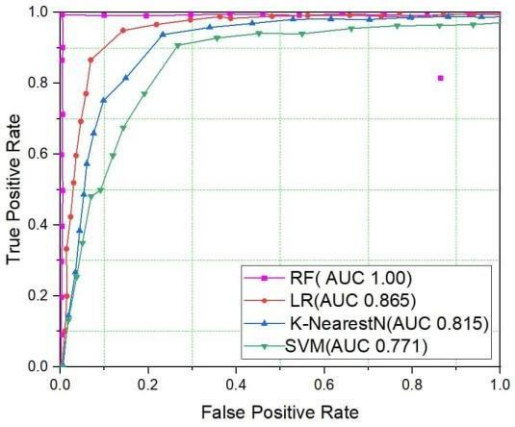

Fig.5. ROC Curve with AUC for PIMA dataset

The cross validation process was also used to assess the efficacy of various models. A subset of the data is set aside for cross validation, and the remaining data is used to train it. And the procedure is repeated for each segment of data. The size of the pieces is determined by the value of k. Here for validation point of view 10-fold cross validation was used, which means the data was split into ten parts. Cross validation has the greatest accuracy for random forest. Random forest has a Kappa statistic of better than 0.9, indicating that it is outstanding.

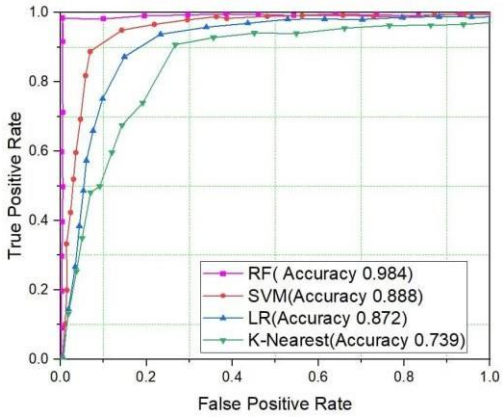

Fig.6. ROC Curve with AUC for PIMA dataset

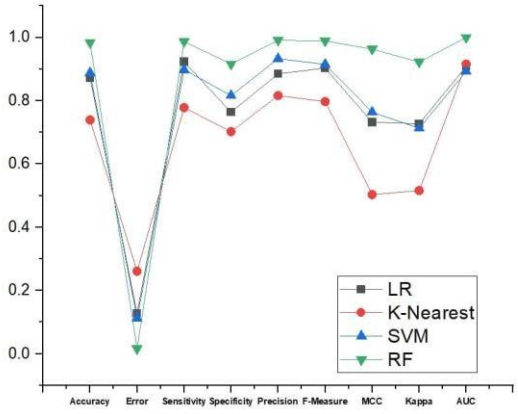

Fig.7. Comparison of different classification approaches

CONCLUSION

One of the most pressing worldwide health concerns is detecting diabetes risk at an early stage. Here in our research that aims to build up a system for predicting the risk of diabetes mellitus. Four machine learning techniques for classification algorithms were used in this work, and the results were compared to several statistical metrics. The above said four algorithms were used on the PIMA database. The accuracy level of RF classification in our dataset is 98.4 percent, which is the greatest among the others, according to the testing results. All of the models generated good results for various parameters like as accuracy, recall sensitivity, and so on, using four different machine learning methods. This result can be used to forecast any other illness in the future. This study is currently researching and improving on various machine learning approaches to forecast diabetes or any other condition.

.

REFERENCES

[1] Qin, Hailun, et al. "Triglyceride to high-density lipoprotein cholesterol ratio is associated with incident diabetes in men: A retrospective study of Chinese individuals." Journal of Diabetes Investigation 11.1 (2020): 192-198. [DOI:10.1111/jdi.13087]

[2] Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299. [DOI:10.1016/j.procs.2020.01.047]

[3] Dash S., Gantayat P.K., Das R.K. (2021) Blockchain Technology in Healthcare: Opportunities and Challenges. In: Panda S.K., Jena A.K., Swain S.K., Satapathy S.C. (eds) Blockchain Technology: Applications and Challenges. Intelligent Systems Reference Library, vol 203. Springer, Cham. https://doi.org/10.1007/978-3-030-69395-4_6

[4] Swapna, G., Vinayakumar R., Soman K. P. (2018) "Diabetes detection using deep learning algorithms." ICT Express 4 (4): 243-246.

[5] Md. Faisal Faruque, Asaduzzaman and I. H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1-4, doi: 10.1109/ECACE.2019.8679365.Dd

[6] Jyotismita Chaki, S. Thillai Ganesh, S.K Cidham, S. Ananda Theertan, Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review, Journal of King Saud University - Computer and Information Sciences, 2020, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2020.06.013Arg

[7] Choi, B.G., Rha, S. W., Kim, S. W., Kang, J. H., Park, J. Y., Noh, Y. K. (2019) "Machine Learning for the Prediction of New-Onset Diabetes Mellitus during 5-Year Follow-up in Non-Diabetic Patients with Cardiovascular Risks." Yonsei medical journal 60 (2): 191-9.Aerg

[8] K. VijiyaKumar, B. Lavanya, I. Nirmala and S. S. Caroline, "Random Forest Algorithm for the Prediction of Diabetes," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2019, pp. 1-5, doi: 10.1109/ICSCAN.2019.8878802.

[9] N. Nnamoko, A. Hussain and D. England, "Predicting Diabetes Onset: An Ensemble Supervised Learning Approach," 2018 IEEE Congress on Evolutionary Computation (CEC), 2018, pp. 1-7, doi: 10.1109/CEC.2018.8477663.

[10] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13, DOI: 10.9790/9622-0801020913

[11] Sisodia, D., Sisodia, D. S. (2018) "Prediction of diabetes using classification algorithms." Procedia computer science 132: 1578-1585.

[12] Sachikanta Dash, Pradosh Kumar Gantayat, 2020 "Liver Disease Prediction Using Machine Learning Algorithm" , Data Engineering and Intelligent Computing, Proceedings of ICICC 2020, https://link.springer.com/chapter, https://doi.org/10.1007/978-981-16-0171-2

[13] Eswari, T., Sampath, P., Lavanya, S. (2015) "Predictive methodology for diabetic data analysis in big data." Procedia Computer Science 50: 203-208.

[14] Perveen, S., Shahbaz, M., Keshavjee, K., Guergachi, A. (2019) "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques." IEEE Access 7: 1365-1375.

[15] https://www.kaggle.com/uciml/pima-indians-diabetes-database