Prognosis of Diabetes Mellitus Based on Machine Learning Algorithms

Ayasha Malik
Department of CSE
Delhi Technical Campus (DTC), GGSIPU
Greater Noida, India
ayasha07.am@gmail.com

Veena Parihar
Department of CSE(AI)
KIET Group of Institutions
Delhi-NCR, Ghaziabad, India
veena2parihar@gmail.com

Jaya Srivastava
Department of CSE
ABES Engineering College
Ghaziabad, India
jayacs0013@gmail.com

Harpreet Kaur
Department of Applied Science
Delhi Technical Campus (DTC), GGSIPU
Greater Noida, India
hkarora92@gmail.com

Shafiqul Abidin
Department of Computer Science
Aligarh Muslim University
Aligarh, Uttar Pradesh, India
shafiqulabidin@yahoo.co.in

***Abstract*— One of the life-threatening and deep root diseases leading to raise the level of sugar in the blood is Diabetes Mellitus (DM) and if it is kept anonymous and untouched then many difficulties have to be faced as a result. Generally, when it is identified a patient visits a diagnostic center and takes a consultation with the doctor. Now, this critical problem can be solved using Machine Learning (ML) approaches as it is booming. A model has been designed in this study that can prognosticate the prediction of DM in patients with a certain level of accuracy. Hence 3 ML algorithms based on classification namely Decision tree, Support Vector Machine (SVM), and Naïve Bayes are used in our study for early detection of DM. Database which is sourced from UCI ML repository namely Pima Indians Diabetes Database (PIDD) is used for performing experiments. Measurement of the performance of all 3 algorithms is based on various criteria like Accuracy, Recall, Precision, and F- Measure. Measurement of Accuracy is done on both correct and incorrect classified instances. After comparative analysis, it has been found that Naïve Bayes is having the highest accuracy among other algorithms. For verification of results Receiver Operating Characteristic (ROC) curves are traced in an organized manner.**

Keywords*— *Healthcare, Diabetes Mellitus, PIMA, Machine Learning, Patients, Decision Tree, Support Vector Machine, Naïve Bayes

I. INTRODUCTION

In the field of medicine, it is needed to categorize the data set in so many classes, therefore many different classification techniques are used in this regard which are based on some constraints rather than a single classifier. The production of Insulin hormone is affected and the body is unable to produce the same or not to the required level for proper functioning in the Diabetes Mellitus (DM) disease. Due to this level of glucose in the blood is increased also there is an abnormality in the metabolism of carbohydrates. DM leads to high blood sugar in the person suffering from it. Some of the frequent symptoms that can be seen in DM are escalated thirst and hunger pangs and the urge to urinate repeatedly. If DM remains untreated it may lead to severe issues later. Some of them are non-ketonic hyperosmolar coma and diabetic ketoacidosis. DM is having a serious concern as there is no control over the measure of sugar level in the blood. Though height, insulin, weight, and hereditary factor are the factors that lead to DM the main reason is the high amount of sugar in the blood. To be safe from the complications of DM there is a need for early detection of the disease. That's why many researchers are using various algorithms based on classification under Machine Learning (ML) for doing experiments for detecting the disease [1]. The work of many researchers is explained, Kalagotla et al. [2] described the PIMA Indian Diabetes (PID) database that was gained from the University/Irvine UCI ML warehouse for investigation commitments. The research was conducted in three phases named, the merging process based on the selection of features, the AdaBoost process was applied to the selected components of the classification, and a fresh installation process with Multiple Layers of Perceptron (MLP), SVM, and Logistic Regression (LR) was planned and established for the designated structures. Tigga et al. [3] measured the insecurity of DM in the person based on their health, way of living, and heredity. The threat of type 2 DM was expected via dissimilar ML methods as per extremely sophisticated algorithms that are highly sought in the medical sector. When the proposed model will be trained with decent accurateness then the people can assess the risk or threat of DM properly. Zhu et al. [4] suggested a model based on the mining of data that was designed for the initial diagnosis and prediction of DM using the PID database. The main objective was to find the

techniques used to develop the K-means clustering and LR effects. The proposed model consists of Principal Components Analysis (PCA), K-means, and LR methods. Anwar et al. [5] implemented a literature survey on curing methods of DM using Artificial Intelligence (AI), Neural Networks (NN), ML, hybrid methods, deep learning, clustering, and data mining. The main aim of this study is to highlight all the limitations of surviving mechanisms that prevent them from providing better curing of DM. Khanam et al. [6] used the PID database, composed of the UCI ML repository database that comprehends all related statistics of 768 patients along with their 9 distinct symptoms. Seven ML algorithms are used to predict DM, it is found that the LR and SVM inherit models work very well in DM prognosis. Moreover, the NN model proposed with an unlike hidden layer with different times, and it is noted that two unseen layers gave an accuracy of 88.6%. Vaishali et al. [7] improved the accurateness of present diagnostic techniques for calculating type 2 DM along with ML algorithms. The suggested algorithm chooses the key structures in the PID dataset collaborates with Goldberg's Genetic (GG) algorithm in the pre-treating phase and uses the MOEF classifier in the database. Hence, the total number of structures was reduced to 4 from 8, and the separation rate was enhanced to 83.0435%. Gupta et al. [8] used sklearn to generate a PID database model and equated many algorithms or methods to gain the finest efficiency. Expecting or guessing DM in females is more important as it not only confirms the start of cure in the initial phase and also aids to protect against highly developing conditions. By calculating many algorithms, it is discovered, which area desires to be worked on to improve better health care methods. Shahriare et al. [9] described a hybrid ML expecting a model that can easily identify type 2 DM even more professionally as compared to preceding activities by gathering PID data from Kaggle UCI ML. after that, outliners are classified and spotted by looking at the interquartile series from this database. A small sampling process was used to estimate this data. After that, the diabetic data was sorted using a simple combination of k-mean clustering. Ahmad et al. [10] compared the correctness of the multilayer perceptron prediction at NN against tree-based algorithms, particularly ID3 and J48 algorithms in the PID mellitus data set, and determined DM or non-DM category and data of 768 patients. The outcomes exposed that the J48 tree algorithm was made an extraordinary exactness that is 89.3% as equated to the multi-layer perceptron algorithm that is 81.9%. Temurtas et al. [11] comprehended a comparative understanding of PID. For this, a multilayer NN model was developed and the NN framework was also used. The outcomes of the training were compared with the results of prior pieces of training. It is reportedly focused on DM testing the same UCI ML database. Huang et al. [12] identified the key factors that contribute to DM management, through the use of feature extraction methods and data mining methods in an effective patient administration organization to support better quality, grouping, and discovery of information. To better develop performance and the overall status of the proposed system, a feature selection procedure by using supervised-based construction was used along with improved ReliefF. The key motivation of this work is as follows.

- The work highlights the various categories of DM with all facts ad figures for a better understanding of the disease so that better and improved curing will be provided to the patients.
- The work deliberates the numerous pre-processing techniques along with detail of various related studies in this domain.
- The work described and implement the ML techniques named Naïve Bayes, SVM, and, Decision tree.
- The work implements the proposed model based on ML techniques for better identification, classification, and diagnosis of DM.
- Finally, the work concludes the result after testing the model based on some decision parameters.

The rest of the paper is organized as follows, section 2 provides the summarizing presentation of DM along with its categories named pre-DM, Type-1 DM, and Type-2 DM. Moreover, section 3 provides reviews or work of some researchers in this era, where many already proposed systems or models are discussed. Furthermore, section 4 elaborates on the methodology of the proposed model along with a working procedure in the form of a flowchart. Additionally, the paper defines the three used ML techniques named Naïve Bayes, SVM, and Decision tree are explained with their calculated confusion matrices based on the dataset obtained from PIDD. Further, some decision parameters are explained to find the efficiency of the proposed model. Section 5 presents the obtained results in the form of graphs. Finally, the paper concludes with section 6

II. Diabetes Mellitus

The theory of DM involves the process of conversion of sugar namely glucose by the body when we have food and transferring it to the blood vessels. Insulin is produced by the pancreas, a hormone whose work is to pass on glucose from the blood to each cell of the body and later use this as energy[35-36]. The body is unable to produce insulin if a person is suffering from DM and not taking medicines and it may cause serious health issues. It depends on what the source is for the development of DM in a person.

A. Pre-Diabetes Mellitus

Whenever blood sugar levels increase to the level that they should be and are not much strong for the doctors to acknowledge it as DM then there is an occurrence of Pre-DM.

There is an increase in Type-II DM and heart disease with the possibility of pre-DM. For reducing these dangers there is a need to reduce weight by 7.3%-9.6% of the body weight and exercise too.

B. Type-1 Diabetes Mellitus

DM based on insulin is called Type-1 DM and is also referred to as juvenile-onset DM, in this category the young people are mostly targeted. An autoimmune condition is there with DM type-1 in which the organs are weakened and lost the ability to generate insulin in the human body. These are some serious health issues that occur due to type-1 DM, disruption to narrow blood vessels in the kidneys (Diabetic Nephropathy), eyes (Diabetic Retinopathy), stroke, and heart failure.

C. Type-2 Diabetes Mellitus

This category of DM is referred to as adult-onset DM and insulin-independent DM. For the last 25 years, it has been widespread among children and teenagers, due to their increasing weight. There is some kind of insulin secretion from the pancreas while you are having type-2 DM [33]. But it isn't sufficient and the human body is unable to consume it as it would. If we compare then Type-2 DM is somewhat milder than Type-1 DM. But it can also cause remarkable health problems such as the complications in the small blood vessels in the nerves, eyes, and kidneys as well as chances of stroke and heart failure also get increased because of this [13].

III. REVIEW OF PRE-PROCESSING TECHNIQUES OF RELATED WORK

Dogantekin et al. [14] offered a comprehensive DM screening program for LDA and the ANFI system named LDA-ANFIS. The arrangement of this intellectual LDA-ANFIS DM indicative system is made up of two stages namely the LDA stage and grouping using the ANFIS classification stage. Jarullah et al. [15] used a decision-making drug approach to predict patients with progressive DM. The database used is the PID database that gathers data of both patients having DM and those without diabetics. In the pre-processing stage, data identification, selection, management of lost values, and optional pricing are done. Moreover, the development of a diabetic prediction model using the decision-making drug approach is done. Patil et al. [16] proposed a hybrid prediction model that uses a simple K-mean clustering algorithm to verify selected labels of data extracted from PID that randomly separated, removed conditions, and later uses a split algorithm in the result dataset. The C4.5 algorithm is used to construct the final separation model using the K-fold cross-verification scheme. Kandhasamy et al. [17] equated the effectiveness of those algorithms that are used to calculate DM by using data mining methods like ML classifiers (J48, Decision Tree (DT), Random Forest (RF), SVM) to distinguish diabetic patients by taking data from the PID UCI ML database. The effectiveness of the algorithms is measured before and after the processing of data and analyzed on the basis of accuracy, sensitivity, and specificity. Han et al. [18] used SVM to diagnose DM and included a learning ensemble segment that converts the "black box" of SVM conclusions into clear and explicit rules, it is also useful in resolving the problem related to inequality. As a result, the proposed production learning model forms sets of rules with an average accuracy of 94.2% and an average rating of 93.9% in all classes. Ganji et al. [19] used an Ant colony-based separation system to produce a set of non-diabetic diagnostic rules, called FCS-ANTMINER, with new features that made it different from surviving methods that use Ant Colony Optimisation (ACO) for grouping of functions. The accurateness of the grouping obtained is 84.24%. Varma et al. [20] have established a tree-cutting model to calculate the happening of DM as traditional decision tree methods have a problem with pleasant boundaries. A key step used in the creation of the decision tree is the identification of the dividing points using the Gini index. In addition, a method is proposed in a way to reduce the prediction of Gini indices by recognizing false points. Nai-arun et al. [21] differentiated the risk of DM. Four known species in the classification model are DT, Artificial NN, LR, and Naive Bayes (NB) were first tested. Subsequently, bagging and boosting approaches were inspected to expand the durability of such models. Moreover, RF was used to estimate this study. Furthermore, the proposed model was used to generate a web-based application to predict the risk category of DM. Maniruzzaman et al. [22] Gaussian Process Classification (GPC), reasonable classification of the diabetic dataset, data examination by cross-verification method and clarification of examined data, and marking of the prescribed technique are discussed using LDA, Quadratic Discriminant Analysis (QDA), classification based on GPC and NB process. Guo et al. [23] proposed the Bayes network expect patients with type 2 DM. The database used is the PID database that assembles the related data of patients with type 2 DM and beyond type 2 DM, exact results were gained. Schizas et al. [24] provided an understanding of how best is to maximize the success rates of segregation achieved through the use of common methods such as NN, RF networks. As a result, the first compatible subsection is used to accomplish the SVM with RF kernel and the second collected subsection is used to accomplish another SVM with a polynomial kernel. Kahramanli et al. [25] proposed a method that accomplished accurate values of 84.24% and 86.8% on behalf of the PID DM database and the database of Cleveland heart disease, correspondingly. These results were found to be one of the best results equated to the results achieved from preceding correlated revisions and testified on UCI websites. Seera et al. [26] suggested a smart hybrid system consisting of fuzzy min-max NN, classification, regression tree, and RF model. It can read increasingly from the data model, interpret its predicted results, and reach higher classification performance. Howlader et al. [27] projected the strictness of DM and the acquisition of important related factors. Records of diabetic patients are collected from the Noakhali Diabetes Association (NDA), Noakhali, Bangladesh. Therefore, raw data can be processed by substituting and deleting lost or incorrect records. Analysis of the database is based on CDT, J48, NBTree, REPtree, and decision-making strategies. Nilashi et al. [28] categorized DM mellitus by establishing a system using ML methods through clustering, noise exclusion, and separation methods. Similarly, the increasing expectations, analysis of the main component and SVM integration, noise exclusion, PCA, and classification functions have been used.

IV. METHODOLOGY USED

A model diagram representation is shown below in figure 1 to represent the proposed flow of research methodology while developing the model.

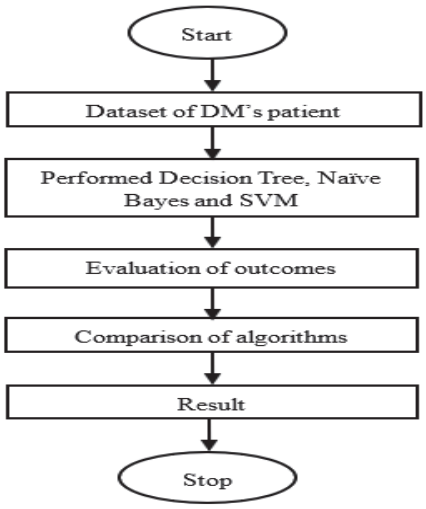

Fig. 1. Flowchart of the proposed model

A. Support Vector Machine (SVM)

Under the supervised ML model, SVM is one of the most important and standard algorithms used for classification [29]. In this algorithm, a small part of two different groups is given for training, the main objective of SVM is to discover the highest level line that divides the hyperplane among the two given groups. Scientifically, the space amongst the hyperplane that is well-defined by $v^t y + c = -1$ and the hyperplane is well-defined by $v^t y + c = 1$ and distance $= \frac{2}{||v||}$. That is we have to solve a maximum of $\frac{2}{||v||}$, consistently, and a minimum of $\frac{||v||}{2}$. The SVM must also appropriately categorize all the y(i), which means $z^i(v^t y^i + c) >= 1$, for all 1 belongs to N. Confusion matrix of the estimated outcome of Naïve Bayes algorithm is shown in table 1.

TABLE I. STATISTICS OF SVM IN FORM OF THE CONFUSION MATRIX

	X	Y
X	536	0
Y	216	0

X Negative
Y Positive

B. Naive Bayes Classifier

Under the classification technique, Naive Bayes classifier is a notion that describes all the self- determining characteristics that are not associated with each other [30]. It explains that the position of a particular characteristic in a group does not impact the position of another feature. It is believed as a strong algorithm that is working for classification purposes, as it is based on Conditional probability. The probability of succeeding targeted group P(M|N) can be estimated from P(M), P(N), and P(N|M) by using the Bayes theorem.

$$P(M|N) = (P(N|M)\ P(M))/P(N)$$

Where,

P(M|N) = probability of succeeding targeted group, P(N|M) = probability of analyst group, P(M) = probability of positive elements of group M and P(N) = preceding probability of analyst. The confusion matrix of the estimated outcome of the Naïve Bayes algorithm is shown in table 2.

TABLE II. STATISTICS OF NAÏVE BAYES IN FORM OF A CONFUSION MATRIX

	X	Y
X	492	61
Y	99	139

X Negative
Y Positive

C. Decision Tree Classifier

Under a supervised ML algorithm, a decision tree is used to solve classification-based problems. The main reason behind using the decision tree in this proposed model is for the accurate prediction of target classes that are using the predefined decision rules taken from already present available data. For the prediction and classification [31] nodes and internodes are used. The confusion matrix of the estimated outcome of the decision tree classifier algorithm is shown in table 3.

TABLE III. STATISTICS OF DECISION TREE IN FORM OF THE CONFUSION MATRIX

	X	Y
X	509	113
Y	116	210

X Negative
Y Positive

D. Dataset Used

An open-source software namely WEKA tool [32] has been used for executing the experiment in this work. It contains a group of several ML approaches for the data classification, clustering, regression, visualization, etc [34]. Predicting the patient if he/she is affected by DM by taking the help of WEKA tool and using the medical database PIDD is the goal of this study. A brief explanation of the dataset is shown in Table-4 and the elements are represented in Table-5.

TABLE IV. REPORT OF USED DATASET

Number of Elements	Number of Occurrences	Database
9	893	PIDD

TABLE V. DESCRIPTION OF ELEMENTS

S. No.	Elements
1	Total number of times women get pregnant
2	Age
3	BP (mm hg)
4	BMI (kg/m2)
5	The thickness of the skin fold
6	The concentration of plasma glucose
7	The pedigree function of DM
8	Insulin
9	Class ‘0’ and ‘1’

In the medical details present in the dataset there are 893 occurrences of female patients. It also includes a numeric value of 8 attributes where the value of one group '0' is considered as tested negative for DM and the value of another group '1' is considered as tested positive for DM.

E. Decision parameters

This research work takes into consideration some algorithms like Naive Bayes, SVM, and Decision Tree. Internal cross-validation 10-folds is used for executing the experiments. For the classification purpose Accuracy, F-Measure, Recall, Precision, and Receiver Operating Curve (ROC) measures are used.

1) Accuracy

Accuracy is defined as 'the amount to which the outcome of a measurement fits the appropriate value or a standard' means how close a measurement is to its agreed value.

Accuracy = (Total number of TP + Total number of TN) / Total number of samples

Where, TP stands for true positive occurrences and TN denotes the true negative occurrences.

2) Recall

A recall is used to measure the classifier’s fullness or sensitivity. Basically, it is the ratio of correct positive predictions to the total positive examples.

Recall = Total number of TP / (Total number of TP + Total number of FN)

Where, TP denotes true positive occurrences and FN denotes the false negative occurrences.

3) Precision

For the model to be 100% precise there should not be any bad positives. It is the ratio of correct positive predictions to total predicted positives.

Precision = Total number of TP / (Total number of TP + Total number of FP)

Where, TP denotes true positive occurrences and FP denotes false positive occurrences.

4) F-measure

F-Measure can be defined as the weighted average of recall and precision. It delivers a way to associate both recall and precision into a single measure that captures both properties. A complete story is not possible with simply a single recall or precision value. A single score is provided to check if we can have a good precision with bad recall or vice versa.

F-measure = 2 (Precision * Recall) / (Precision + Recall)*

5) ROC

For the comparison of the efficacy of tests, ROCs are used. Here the performance of the classification model at all thresholds is shown using a graph.

V. RESULTS

The performance of all classification algorithms is calculated on many decision parameters represented in Table-6. Next, the performance based on classified occurrences is determined in Table-7. Accuracy is calculated and analyzed based on these categorized occurrences. Out of the total number of correct and incorrect occurrences, a performance evaluation is done for each algorithm. The naive Bayes algorithm outclasses as the best algorithm in comparison to other algorithms and is shown in Table-6 and Table-7. Hence, the best-supervised ML algorithm of this experiment came to be a Naïve Bayes algorithm. It gives an accuracy of 86.21% which is way higher than the rest algorithms.

TABLE VI. OUTCOMES OF ALGORITHMS BASED ON DECISION PARAMETERS.

Algorithm	% of Accuracy	Recall	Precision	F-measure	ROC
Decision tree	74.5	0.798	0.785	0.762	0.806
Naïve Bayes	86.21	0.726	0.746	0.736	0.726
SVM	68.56	0.645	0.495	0.535	0.503

TABLE VII. OUTCOMES OF ALGORITHMS BASED ON CORRECT AND INCORRECT OCCURRENCES

Total number of occurrences	Algorithm	Correct occurrences	Incorrect Occurrences
893	Decision tree	701	299
	Naïve Bayes	686	282
	SVM	653	301

Figure 2, figure 3, figure 4, and figure 5 show the performance of all classifiers which are based on many measures and are plotted using a graph. The representation of ROC area of all classification algorithms is shown in Figure 6.

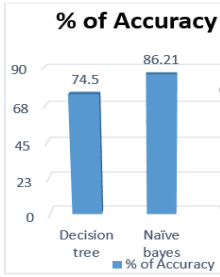

Fig. 2. Graph of percentage of accuracy

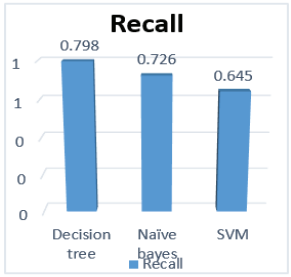

Fig. 3. Graph of recall

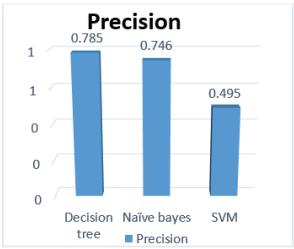

Fig. 4. Graph of precision

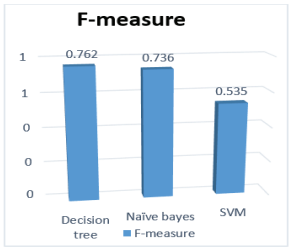

Fig. 5. Graph of F-measure

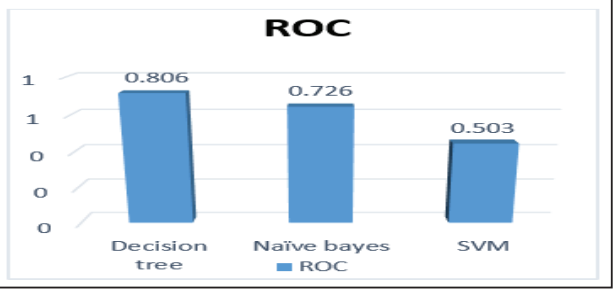

Fig. 6. Graph of ROC

VI. Conclusion

The utmost important difficulty in the physical world is the discovery of DM at its initial phase. This work presented efficient efforts to develop a system that led to the forecast of serious diseases like DM. Furthermore, this paper explored DM with its types. Moreover, the paper delivered information about a few pre-implemented projects along with the methodology of our work. Additionally, the paper described some prior algorithms of ML named Naïve Bayes, SVM, and Decision tree with their evaluated confusion matrices based on a dataset obtained from PIDD, and those algorithms are also tested on the basis of various decision parameters. The research is implemented on the database of Pima Indians and the outcome concludes the competence of the 86.21% by using the Naive Bayes ML algorithm. In future research, the same system could use to forecast and cure other harmful diseases by use of some other ML techniques.

References

[1] Chauhan T, Rawat S, Malik S and Singh P, (2021) Supervised and Unsupervised Machine Learning-based Review on Diabetes Care," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 581-585, doi: 10.1109/ICACCS51430.2021.9442021.

[2] Kalagotla S.K, Gangashetty S.V, Giridhar K, (2021) A novel stacking technique for prediction of diabetes, Computers in Biology and Medicine, Volume 135, 104554, ISSN 0010-4825, doi.org/10.1016/j.compbiomed.2021.104554.

[3] Tigga N.P, Garg S, (2020) Prediction of Type 2 Diabetes using Machine Learning Classification Methods, Procedia Computer Science, Volume 167, 2020, Pages 706-716, ISSN 1877-0509, doi.org/10.1016/j.procs.2020.03.336.

[4] Zhu C, Idemudia C.U, Feng W, (2019) Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques, Informatics in Medicine Unlocked, Volume 17, 100179, ISSN 2352-9148, doi.org/10.1016/j.imu.2019.100179.

[5] Anwar F, Ain Q.U, Ejaz M.Y, Mosavi A, (2020) A comparative analysis on diagnosis of diabetes mellitus using different approaches – A survey, Informatics in Medicine Unlocked, Volume 21, 100482, ISSN 2352-9148, doi.org/10.1016/j.imu.2020.100482.

[6] Khanam J.J, Foo S.Y, (2021) A comparison of machine learning algorithms for diabetes prediction, ICT Express, ISSN 2405-9595, doi.org/10.1016/j.icte.2021.02.004.

[7] Vaishali R, Sasikala R, Ramasubbareddy S, Remya S and Nalluri S, (2017) Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset, International Conference on Computing Networking and Informatics (ICCNI), pp. 1-5, doi: 10.1109/ICCNI.2017.8123815.

[8] Gupta A.D., Bhattacharyya S., Snasel V., Platos J., Hassanien A. International Conference on Innovative Computing and Communications, Advances in Intelligent Systems and Computing, vol 1087. Springer, Singapore, doi.org/10.1007/978-981-15-1286-5_29

[9] Shahriare S.M., Atik S.T., Moni M.A. (2020) A Novel Hybrid Machine Learning Model to Predict Diabetes Mellitus. In: Uddin M.S., Bansal J.C. (eds) Proceedings of International Joint Conference on Computational Intelligence. Algorithms for Intelligent Systems, Springer, Singapore, doi.org/10.1007/978-981-15-3607-6_36

[10] Ahmad A., Mustapha A., Zahadi E.D., Masah N., Yahaya N.Y. (2011) Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus. In: Snasel V., Platos J., El-Qawasmeh E. (eds) Digital Information Processing and Communications. ICDIPC 2011. Communications in Computer and Information Science, vol 188, Springer, Berlin, Heidelberg, doi.org/10.1007/978-3-642-22389-1_47

[11] Temurtas H, Yumusak N, Temurtas F, (2009) A comparative study on diabetes disease diagnosis using neural networks, Expert Systems with Applications, Volume 36, Issue 4, 2009, Pages 8610-8615, ISSN 0957-4174, doi.org/10.1016/j.eswa.2008.10.032.

[12] Huang Y, McCullagh P, Black N, Harper R, (2007) Feature selection and classification model construction on type 2 diabetic patients' data, Artificial Intelligence in Medicine, Volume 41, Issue 3, 2007, Pages 251-262, ISSN 0933-3657, doi.org/10.1016/j.artmed.2007.07.002.

[13] Mohan V and Pradeepa R, (2014) Telemedicine in Diabetes Care: In rural India, a new prevention project seeks to fill in the screening gap., in IEEE Pulse, vol. 5, no. 3, pp. 22-25, May-June 2014, doi: 10.1109/MPUL.2014.2309575.

[14] Dogantekin E, Dogantekin A, Avci D, Avci L, (2010) An intelligent diagnosis system for diabetes on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System: LDA-ANFIS, Digital Signal Processing, Volume 20, Issue 4, 2010, Pages 1248-1255, ISSN 1051-2004, doi.org/10.1016/j.dsp.2009.10.021.

[15] Jarullah A.A.A, (2011) Decision tree discovery for the diagnosis of type II diabetes," 2011 International Conference on Innovations in Information Technology, 2011, pp. 303-307, doi: 10.1109/INNOVATIONS.2011.5893838.

[16] Patil B.M, Joshi R.C, Toshniwal D, (2010) Hybrid prediction model for Type-2 diabetic patients, Expert Systems with Applications, Volume 37, Issue 12, 2010, Pages 8102-8108, ISSN 0957-4174, doi.org/10.1016/j.eswa.2010.05.078.

[17] Kandhasamy J.P, Balamurali S, (2015) Performance Analysis of Classifier Models to Predict Diabetes Mellitus, Procedia Computer Science, Volume 47, 2015, Pages 45-51, ISSN 1877-0509, doi.org/10.1016/j.procs.2015.03.182.

[18] Han L, Luo S, Yu J, Pan L and Chen S, (2015) Rule Extraction From Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes," in IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 2, pp. 728-734, March 2015, doi: 10.1109/JBHI.2014.2325615.

[19] Ganji M, Abadeh M.S, (2011) A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis, Expert Systems with Applications, Volume 38, Issue 12, 2011, Pages 14650-14659, ISSN 0957-4174, doi.org/10.1016/j.eswa.2011.05.018.

[20] Varma K, Rao A.A, Lakshmi T.S.M, Rao P.V.N, (2014) A computational intelligence approach for a better diagnosis of diabetic patients, Computers & Electrical Engineering, Volume 40, Issue 5, 2014, Pages 1758-1765, ISSN 0045-7906, doi.org/10.1016/j.compeleceng.2013.07.003.

[21] Nai-arun N, Moungmai R, (2015) Comparison of Classifiers for the Risk of Diabetes Prediction, Procedia Computer Science, Volume 69, 2015, Pages 132-142, ISSN 1877-0509, doi.org/10.1016/j.procs.2015.10.014.

[22] Maniruzzaman M, Kumar N, Abedin M.M, Islam M.S, Suri H.S, El-Baz A, Suri, J.S, (2017) Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm, Computer Methods and Programs in Biomedicine, Volume 152, 2017, Pages 23-34, ISSN 0169-2607, doi.org/10.1016/j.cmpb.2017.09.004.

[23] Guo Y, Bai G and Hu Y, (2012) Using Bayes Network for Prediction of Type-2 diabetes, *2012 International Conference for Internet Technology and Secured Transactions*, 2012, pp. 471-472.

[24] Schizas C.N and Karatsiolis S. (2012), Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset. In Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE) (BIBE '12). IEEE Computer Society, USA, 139–144. doi.org/10.1109/BIBE.2012.6399663

[25] Kahramanli H, Allahverdi N,Design of a hybrid system for the diabetes and heart diseases, Expert Systems with Applications, Volume 35, Issues 1–2, 2008, Pages 82-89, ISSN 0957-4174, doi.org/10.1016/j.eswa.2007.06.004.

[26] Seera M, Lim C.P, (2014) A hybrid intelligent system for medical data classification, Expert Systems with Applications, Volume 41, Issue 5, 2014, Pages 2239-2249, ISSN 0957-4174, doi.org/10.1016/j.eswa.2013.09.022.

[27] Howlader K.C, Satu M.S, Barua A, Moni M, (2018) Mining Significant Features of Diabetes Mellitus Applying Decision Trees: A Case Study In Bangladesh, 2018, doi.org/10.1101/481994

[28] Nilashi M, Bin Ibrahim O, Mardani A, Ahani A, Jusoh A (2018) A soft computing approach for diabetes disease classification. Health Informatics Journal. December 2018:379-393. doi:10.1177/1460458216675500

[29] Borges F et al., (2020) An Unsupervised Method based on Support Vector Machines and Higher-Order Statistics for Mechanical Faults Detection," in IEEE Latin America Transactions, vol. 18, no. 06, pp. 1093-1101, Jun 2020, doi: 10.1109/TLA.2020.9099687.

[30] Subbulakshmi C.V, Deepa S.N and Malathi N, (2012) Comparative analysis of XLMiner and WEKA for pattern classification," 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 2012, pp. 453-457, doi: 10.1109/ICACCCT.2012.6320821.

[31] Shariati, M., Mafipour, M.S., Ghahremani, B. *et al.* (2020) A novel hybrid extreme learning machine–grey wolf optimizer (ELM-GWO) model to predict compressive strength of concrete with partial replacements for cement. *Engineering with Computers* (2020). Doi: 10.1007/s00366-020-01081-0

[32] Akram, M.U., Khan, S.A. (2013) Multilayered thresholding-based blood vessel segmentation for screening of diabetic retinopathy. *Engineering with Computers* **29,** 165–173 (2013) doi: 10.1007/s00366-011-0253-7

[33] A. Sarwar, M. Ali, J. Manhas, and V. Sharma, "Diagnosis of diabetes type-II using hybrid machine learning based ensemble model," *Int. J. Inf. Technol.*, vol. 12, no. 2, pp. 419–428, 2020, doi: 10.1007/s41870-018-0270-5.

[34] Thai, DK., Tu, T.M., Bui, T.Q. *et al.* (2021) Gradient tree boosting machine learning on predicting the failure modes of the RC panels under impact loads. Engineering *with Computers* **37,** 597–608 (2021) doi: 10.1007/s00366-019-00842-w

[35] S. Chakraborty, G. C. Jana, D. Kumari, and A. Swetapadma, "An improved method using supervised learning technique for diabetic retinopathy detection," *Int. J. Inf. Technol.*, vol. 12, no. 2, pp. 473–477, 2020, doi: 10.1007/s41870-019-00318-6.

[36] I. Hasan, P. Dhawan, S. A. M. Rizvi, and S. Dhir, "Data analytics and knowledge management approach for COVID-19 prediction and control," *Int. J. Inf. Technol.*, vol. 15, no. 2, pp. 937–954, 2023, doi: 10.1007/s41870-022-00967-0.