2022 Iraqi International Conference on Communication and Information Technologies (IICCIT) | 978-1-6654-7220-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/IICCIT55816.2022.10010408

Detecting Diabetes Using Machine Learning Algorithms

Qabeela Q. Thabit
Basrah Education Directorate
Ministry of education
Basrah, Iraq
qabeelaqassim@basrahaoe.iq
qabelh2010@gmail.com

Taqwa O. Fahad
Biomedical Engineering Department
University of Technology- Iraq
Baghdad, Iraq
taqwa.o.fahad@uotechnology.edu.iq

Alyaa I. Dawood
Engineering Technical College
Southern Technical University
Basrah, Iraq
alyaa.dawood@stu.edu.iq

***Abstract*— Machine learning has become a significant and promising approach to research and has proven to be an effective technique. It can significantly help in the early diagnosis and prediction of diabetes, which has spread rapidly throughout the world, demanding more modern techniques to detect and diagnose diabetes. In this regard, artificial intelligence (AI) has proven its efficiency in detecting diseases, especially for critical discovery cases depending on a database approved by medical officials. This study estimates the incidence of diabetes using classification techniques that employ a variety of machine learning algorithms, such as Logistic Regression (LR) or so-called Classifier, Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), and Naive Bays (NB) classifiers. These techniques are used to determine the prevalence of diabetes in all methods related to the preprocessing data, including cleaning data, feature selection, scaling data, and splitting of data. Depending on patient electronic health data (Pima Diabetes Database of India), the present work is implemented using Python programming language within the ANACONDA environment under Jupyter (notebook 6.1.4) and Spyder (4.1.5) applications, and the outcomes show its ability to obtain accurate models with excellent results that can predict the presence of diabetes with up to 100% accuracy.**

Keywords*— *Machine learning algorithms, Diabetes detection, Performance metrics, Data preparing.

I. INTRODUCTION

Nowadays, Artificial intelligence is involved in all fields of science and research due to its significant potential to obtain results in an economical and efficient manner. The researchers focused on diagnosing diabetes using neural networks and presented results with this application where Diabetes 2 was detected using deep neural network processing [1]. Swapna et al. [2] also detected the use of deep neural networks trained by machine learning methods. Olisah et al. [3] used a twice-growth deep neural network (2GDNN) model and a supervised learning algorithm to develop a diabetes diagnostic system. After systematic training and testing, the neural network achieved a success rate of 99.57%.

With the growing number of human beings with type 1 and type 2 diabetes, research has continued to employ the neural network to detect diabetes because of the results it shows, which are to a large extent considered in this field [4], [5]. Diabetes frequently causes diabetic retinopathy, which harms the blood vessels in the brain's luminous tissue. It is considered the main cause of visual impairment and blindness being the main reason for diabetics lose their sight. Because of its paramount importance, it was presented in various ways in artificial intelligence, therefore, a large number of researchers focused on the use of neural networks to obtain training for the database that includes retina images [6]-[9]. Most research that include images as data use artificial neural network algorithms as methods for machine learning. Neural networks have multiple uses, for example, in various engineering applications such as computation or designing logic gates [10], in additive manufacturing, and ANN networks as a basis for developing solutions to mining and environmental problems. In contrast, the traditional techniques fail in one way or another [11],[12] and in different medical applications [13], [14]. It is noted that the detection of diabetes has taken an expanded approach to include all branches of artificial intelligence that can be used in this concept [15]-[18]. Binary classification algorithms or so-called logistic regression algorithms were implemented on a database of diabetic patients, including logistic regression, decision tree, support vector machine (SVM), XgBoost, Random Forest (RF), and AdaBoost [19], [20].

In this paper, machine learning by binary classification method is employed to predict the presence of diabetes, using Python and depending on the Pima Indian Diabetes Dataset 768. A variety of machine learning classification algorithms involving the Logistic Regression classifier (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), K Nearest Neighbor (KNN), and Naive Bayes (NB) classifier are used to predict diabetes in patients. These algorithms employ various data preparation techniques (cleaning, feature selection, scaling, in addition to the splitting of the data) to produce the best results when evaluated on the dataset.

II. MATERIALS AND METHOD

A. Work Flow Diagram

The procedure of working includes the following steps [3]:

1. Obtain the database; here the Pima Indian dataset is used.
2. Preprocessing Data, including cleaning data, feature selection, scaling data, and splitting of data.
3. Apply algorithms of classification:
   - LR classifier.
   - SVM classifier.

- DT classifier.
- RF classifier.
- GB classifier.
- KNN classifier.
- and NB classifier.

4. Evaluate performance metrics: Accuracy, Sensitivity, Specificity, F1-score, and ROCAUC Score.
5. Prediction of Diabetes.

B. *Dataset*

The global institute, known as the National Institute of Diabetes and Kidney Disease [21], presented the Indian PIMA dataset. There are 768 items in the dataset; each has eight attributes (features) and one outcome (label or output), including two values: 0 indicates no diabetes case (500) in data, and 1 indicates diabetes presented case (268) in the data as shown in Fig. 1.

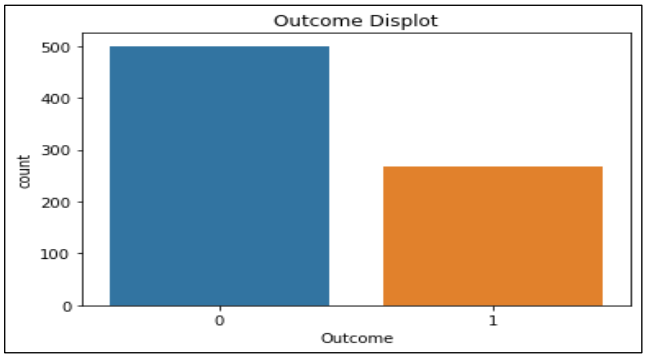

Fig. 1 Outcome of Diabetes.

Based on specific diagnostic criteria offered in the collection, the objective is to detect whether the patient has diabetes or not. There were numerous restrictions that the focus was centered on choosing these cases from a larger database. The collection dataset includes one outcome (dependent) variable as well as a number of medical predictors (independent) components. A patient's age, Body Mass Index (BMI), number of pregnancies, blood pressure in the diastole (mmHg), the thickness of the triceps skin fold (mm), insulin serum during two hours (mu U/ml), diabetes pedigree function and glucose concentration. At this clinic, every patient is a PIMA Indian lady above the age of 21. The extensive features of the databases in PIMA India are outlined in TABLE I for only five rows.

TABLE I. FEATURES AND OUTCOME OF PIMA DATASET.

No.	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.6	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

C. *Preprocessing Dataset*

After the data acquisition stage, the data initialization stage, and this stage includes several points, namely:

- Cleaning data: Sometimes, there are empty cells or missing data within the data, which are processed within a set of instructions to make sure there are no missing data, as shown in TABLE II.

TABLE II. REPORT OF PIMA DATASET.

Features	Value
Pregnancies	0
Glucose	0
Blood Pressure	0
Skin Thickness	0
Insulin	0
BMI	0
Diabetes Pedigree Function	0
Age	0
Outcome	0
dtype : int64	

After that, the complete information was studied in terms of the count and the arithmetic average or mean for each feature and other information representing all the statistical measures, as shown in TABLE III.

TABLE III. DATASET STATISTICAL MEASURES.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845	3.369	0.00	1.00	3.00	6.00	17.00
Glucose	768.0	120.894	31.972	0.00	99.00	117.0	140.25	199.0
Blood Pressure	768.0	69.105	19.355	0.00	62.00	72.00	80.00	122.0
Skin Thickness	768.0	20.536	15.952	0.00	0.00	23.00	32.00	99.00
Insulin	768.0	79.799	115.24	0.00	0.00	30.50	127.25	846.0
BMI	768.0	31.992	7.88	0.00	27.30	32.00	36.600	67.10
Diabetes Pedigree Function	768.0	0.472	0.331	0.078	0.243	0.372	0.626	2.42
Age	768.0	33.241	11.76	21.00	24.00	29.00	41.00	81.00
Outcome	768.0	0.349	0.477	0.00	0.00	0.00	1.00	1.00

- Feature selection: To choose the characteristics affecting the measurements, several methods are used to help obtain better results. In order to find the influential features, a relationship was made between the features, as shown in Fig. 2.

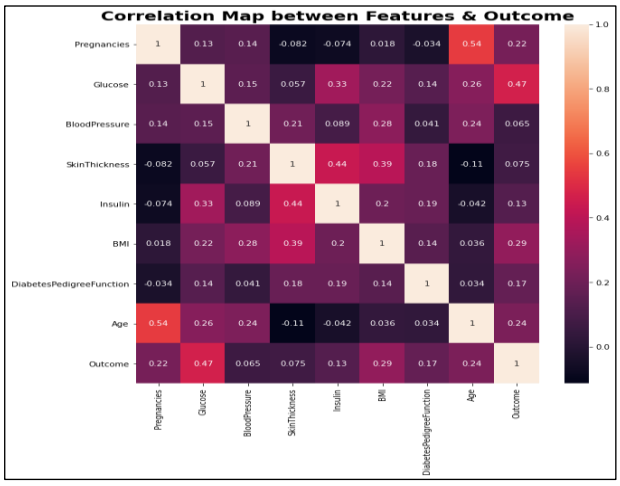

Fig. 2 Correlation between Features and Outcome.

- Scaling data: Attempting to approximate all data on a close digital scale and promotes obtaining correct, accurate, and also fast measurements because the data spacing causes a delay in obtaining the required measures.
- Splitting of data: It is to separate the data into two parts, the training section, which is usually 6-8 %, and the other section is the test section, usually 2-4 %.

III. Classification Methods

Machine learning is an important artificial intelligence branch that is becoming increasingly popular as a means of providing novel services to consumers. Machine learning can be defined as the science that enables a computer to predict and make suitable, effective, and rapid judgments "automatically" through employing algorithms that allow it to do so without the need for any prior knowledge or program experiences. Three sections can be included in machine learning: supervised learning, unsupervised learning, and reinforcement learning. This work dealt with supervised learning algorithms specifically classification algorithms as the following [22]:

A. Logistic Regression

LR is a sort of supervised learning that uses the sigmoid function to evaluate probabilities and estimate the relationship between dependent pairs of variables in addition to at least one individual variable. Contrary to its call, logistic regression represents a sort of machine learning which can be used to solve regression problems. A classification issue with a dichotomous dependent variable (0/1, or -1/1, or represented in true/false) and a binary independent or free variable. The level may be binary, ordinal, interval, or ratio. [23] gives a rule of the sigmoid/logistical function in "(1)" :

$$Y = \frac{1}{1 + e^{-x}} \qquad (1)$$

B. Support Vector Machine

SVM is considered a supervised classifier used as a machine learning technique for prediction regression in machine learning or classification. It is generally used to solve classification situations [24]. SVM's value lies in its ability to classify data points in a multidimensional space using the proper super level. The hyper-level serves as the classification decision boundary for data points. The hyper-plane with the biggest difference between classes and hyper-plane were utilized in order to classify data points. The classification of the supporting vector machine [25] is shown in Fig. 3.

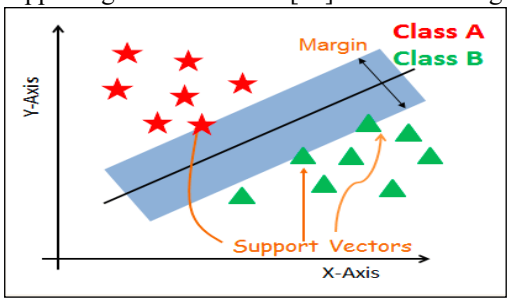

Fig. 3 Super Vector Machine [23].

C. Decision Tree

Another example of a tree-like supervised learning algorithm is a decision tree technique. DTs are usually simple to understand because they simulate people's expectations or thoughts when making decisions. The dataset includes the attributes and features, which will be represented by the inner node, and the branches outside of it will represent the decision base which, in turn, will branch, and the result is tracked by the path that goes to the leaf node. As depicted in Fig. 4, the decision tree makes decisions via splitting each node into sub-nodes, beginning with the first node. Throughout the training phase, this process is repeated numerous times until only homogeneous nodes are ignored. The decision tree only succeeds so well because of this [26], [27].

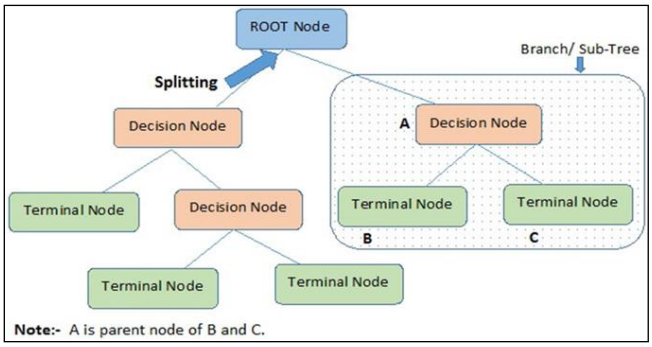

Fig. 4 Decision Tree [28].

D. Random Forest

The set of decision trees constructs a random forest classifier, which is trained through this set. Then the classifier collects the votes of different decision trees in order to determine the final class that is the best solution, as explained in the following Fig. 5 [29].

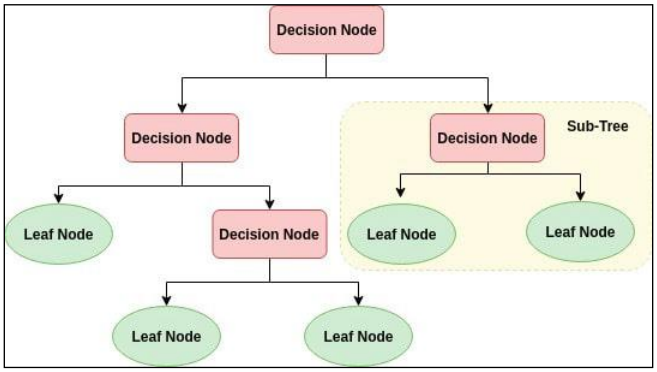

Fig. 5 Random Forest [23].

E. Gradient Boosting

For regression and classification, an ensemble technique is used. Although GB is comparable to the other boosting methods, it can only be used for regression. As illustrated in Fig. 6, each iteration of the random manner selects a training set, which is tested versus the base model of the gradient boosting process. GB for execution can be made faster and more accurate through random sub-sampling of data training. This also helps to avoid over-fitting [30],[31].

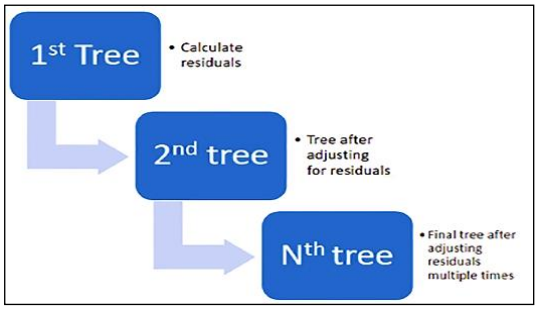

Fig. 6 Gradient Boosting Flow Chart [30].

F. K-Nearest Neighbor

The KNN technique determines the nearest neighbor based on k value, which establishes the number of nearest neighbors to consider when defining the class of a sample data point [21],[23]. Structure-based methodologies cover the underlying structure of the data with fewer mechanisms associated with training data samples. The total data is divided into data points besides training data with a less formal structure, and the distance between the sample points into all other training points is calculated; the point with the smallest distance is known as the closest neighbor, as shown in Fig. 7.

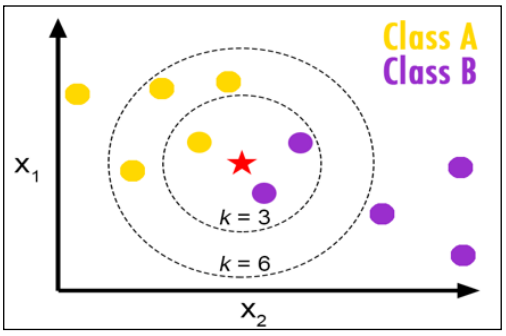

Fig. 7 K-Nearest Neighbor [21].

G. Naïve Bays

The NB classifier is a classification approach that uses Bayes' theorem to maximize subsequent probability. It may also be used to compute a given sample probability depending on their previous probabilities and to improve classification probability in variable and conditional components. Despite its simplicity, this technique produces better results than more complicated classifiers [32]. Thomas Bayes is an English scientist who claims to have invented the probability and statistics-based categorization method. The well-known Bayesian hypothesis of the future [33] predicts probability based on previous experiences.

IV. Result and Discussion

A. Performance Evaluation Metrics

Using the confusion matrix presented in Table I, the performance of the suggested strategy was evaluated. Following are the four possible value outcomes of the confusion matrix:

- True Positive (TP).
- True Negative (TN).
- False Positive (FP).
- And False Negative (FN).

TABLE IV. CONFUSION MATRIX PARAMETERS.

		Prediction case	
		1	*0*
Actual Case	*1*	True positive (TP)	False negative (FN)
	0	False Positive (FP)	True Negative (TN)

Equations 2, 3, 4, and 5 in TABLE V represent the accuracy metric used to compute the performance evaluation parameters for the two classification approaches.

TABLE V. EVALUATION PARAMETERS.

No.	Metric	Definition
1	**Accuracy-Score**	$\frac{TP + TN}{TP + FP + FN + TN}$ (2)
2	**Sensitivity (Recall-Score)**	$\frac{TP}{TP + FN}$ (3)
3	**Specificity (Precision Score)**	$\frac{TP}{TP + FP}$ (4)
4	**F1-Score**	$\frac{2\ (recall \times precision)}{(recall + precision)}$ (5)
5	**ROCAUC Score**	ROC (Receiver Operating Curve) and AUC (Area Under Curve)

In this work, extracted performance measures, such as accuracy, sensitivity, specificity, F1-Score, and ROCAUC-Score for several classifiers were compared with the performance measures for the Indian PIMA Diabetes Database using several classifiers as shown in TABLE VI.

TABLE VI. CLASSIFIERS PERFORMANCE METRICS.

Evaluation Metrics	Accuracy Score	Recall Score	Precision Score	F1 Score	ROCAUC Score
LR	0.9816	0.9816	0.9816	0.9816	0.9714
SVM	1.0	1.0	1.0	1.0	1.0
DT	1.0	1.0	1.0	1.0	1.0
RF	0.9411	0.9411	0.9411	0.9411	0.9411
GB	0.941	0.941	0.941	0.9411	0.954
KNN	0.9	0.9	0.9	0.9	0.9
NB	0.8974	0.8974	0.8974	0.8974	0.8845

In implementing the above algorithms, it was found that the two algorithms SVM and DT are more efficient in predicting diabetes as a higher percentage of performance measures. This leads to directing efforts in this field of scientific research.

B. Comparison of Performance Measures

Seven different classifiers are used in this paper to determine the occurrence of diabetes in the PIMA dataset. It is compared to the previous researcher's work and the results are recorded in Tables VII, VIII, IX, X, XI, XII, and XIII.

TABLE VII. LOGISTIC REGRESSION.

Ref. No.	Authors	Accuracy (%)
[20]	Henock M. Deberneh	0.71
[23]	Neha Prerna T.	0.744
[35]	Nazin Ahmed	0.7763
[34]	Jobeda Jamal Khanam	0.7885
[37]	Tawfik Beghriche	0.790

[19]	Boshra Farajollahi	0.8311
[38]	Changsheng Zhua	0.89

TABLE VIII. SUPPORT VECTOR MACHINE.

Ref. No.	Authors	Accuracy (%)
[38]	Changsheng Zhua	0.58
[36]	Deepti Sisodiaa	0.6510
[21]	Vandana C. Bavkar	0.7386
[23]	Neha Prerna T.	0.744
[34]	Jobeda Jamal Khanam	0.7771
[28]	N. Sneha	0.7773
[35]	Nazin Ahmed	0.8026
[19]	Boshra Farajollahi	0.8246
[37]	Tawfik Beghriche	0.9675

TABLE IX. DECISION TREE.

Ref. No.	Authors	Accuracy (%)
[23]	Neha Prerna T.	0.697
[28]	N. Sneha	0.7318
[36]	Deepti Sisodiaa	0.7382
[34]	Jobeda Jamal Khanam	0.7424
[35]	Nazin Ahmed	0.7632
[19]	Boshra Farajollahi	0.7987
[39]	Seyede Somayeh	0.802
[21]	Vandana C. Bavkar	0.8997
[37]	Tawfik Beghriche	0.9875

TABLE X. RANDOM FOREST.

Ref. No.	Authors	Accuracy (%)
[20]	Henock M. Deberneh	0.73
[23]	Neha Prerna T.	0.750
[28]	N. Sneha	0.7539
[34]	Jobeda Jamal Khanam	0.7714
[35]	Nazin Ahmed	0.8026
[19]	Boshra Farajollahi	0.8311
[41]	Asmita Singh	0.947
[37]	Tawfik Beghriche	0.9920

TABLE XI. GRADIENT BOOSTING.

Ref. No.	Authors	Accuracy (%)
[20]	Henock M. Deberneh	0.72
[19]	Boshra Farajollahi	0.7792
[35]	Nazin Ahmed	0.7895
[38]	Changsheng Zhua	0.85
[37]	Tawfik Beghriche	0.9725

TABLE XII. K-NEAREST NEIGHBOR.

Ref. No.	Authors	Accuracy (%)
[28]	N. Sneha	0.6304
[21]	Vandana C. Bavkar	0.7035
[23]	Neha Prerna T.	0.708
[35]	Nazin Ahmed	0.75
[40]	Muhammad Azeem Sarwar	0.77
[38]	Changsheng Zhua	0.78
[34]	Jobeda Jamal Khanam	0.7942
[41]	Asmita Singh	0.933

TABLE XIII. NAÏVE BAYS.

Ref. No.	Authors	Accuracy (%)
[23]	Neha Prerna T.	0.689
[28]	N. Sneha	0.7348
[40]	Muhammad Azeem Sarwar	0.74
[36]	Deepti Sisodiaa	0.7630
[34]	Jobeda Jamal Khanam	0.7828
[21]	Vandana C. Bavkar	0.7857
[35]	Nazin Ahmed	0.7895
[38]	Changsheng Zhua	0.82

V. CONCLUSION

The field of artificial intelligence, especially machine learning, is currently considered a powerful tool that has impacted all scientific and practical disciplines as it overlaps with all applied sciences. It can be adopted in health care for pathological analysis and changing the traditional method of doctors and practitioners in the medical field. This work is used to predict the disease Diabetes Seven Machine Learning Algorithms. These algorithms are LR, SVC, DT, RF, KNN, and NB classifiers. Diabetes predictions were developed using the Indian PIMA dataset, which contains 768 records. The prediction model was trained and tested using 8 traits. The experimental results reveal that all the employed methods are suitable for predicting diabetes mellitus; among them are SVM and DT, which can achieve the highest accuracy with nearly 100% for diabetes prediction. For future works, it is possible to rely on these two algorithms and link them to transfer or deep learning in order to search for the optimal algorithm. Furthermore, it is possible to rely on a larger database using the same application to show results and compare them with the current ones.

REFERENCES

[1] K. Kannadasan, Edla D. R., and V. Kuppili ., “Type 2 diabetes data classification using stacked auto-encoders in deep neural networks,” Clinical Epidemiology and Global Health, vol. 7, pp.530-535, 2019.

[2] G. Swapna, R. Vinayakumar, and S. Kp, “Diabetes detection using deep learning algorithms,” ICT Express, vol. 4, pp. 243–246, 2018.

[3] C. C. Olisah, L. Smith, and M. Smith, “Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective,” Computer Methods and Programs in Biomedicine, vol. 220, pp. 1-12, 2022.

[4] S. HoLing, PhyoPhyoSan, and Hung T. Nguyen, “Non-invasive hypoglycemia monitoring system using extreme learning machine for Type 1 diabetes,” ISA Transactions.

[5] O. I. Khristodulo, A. A. Makhmutov, and T. V. Sazonova, “Use algorithm based at Hamming neural network method for natural objects classification,” Procedia Computer Science, vol. 103, pp. 388-395, 2017.

[6] A. Pak, A. Ziyaden, K. Tukeshev, A. Jaxylykova, and D. Abdullina, “Comparative analysis of deep learning methods of detection of diabetic retinopathy,” Cogent Engineering, vol. 7, pp.1-9, 2020.

[7] P. Kaur, S. Chatterjee, and D. Singh, “Neural network technique for diabetic retinopathy detection,” International Journal of Engineering

and Advanced Technology (IJEAT), vol.8 no.6, pp, 440-445, August 2019.

[8] S. I Pao et. al, "Detection of diabetic retinopathy using bi-channel convolutional neural network," Journal of Ophthalmology, vol. 2020, pp.1- 7, June 2020.

[9] M. Mohsin Butt, G.r Latif, D. N. F. Awang Iskandar, Jaafar Alghazo, and Adil H. Khan, "Multi-channel convolutions neural network based diabetic retinopathy detection from fundus images," 16th International Learning & Technology Conference 2019 (Procedia Computer Science), vol. 163 , pp. 283–291, 2019.

[10] Q. Q. Thabit, Alyaa Ibragim Dawood, and Bayadir A. Issa, "Implementation three-step algorithm based on signed digit number system by using neural network," Indonesian Journal of Electrical Engineering and Computer Science, vol. 24, No. 3, pp. 1832-1839, December 2021.

[11] M. Valizadeh, S. J. Wolff, "Convolutional neural network applications in additive manufacturing: a review," Advances in Industrial and Manufacturing Engineering, vol. 4, pp.1-12 ,2022.

[12] W. Ma, Y. Du , X.i Liu, and Y. Shen, "Literature review: Multi-criteria decision-making method application for sustainable deep-sea mining transport plans," Ecological Indicators, vol.140, pp. 1-15, 2022.

[13] A. Almarzouqi, A. Aburayya, Said A. Salloum, "Determinants of intention to use medical smartwatch-based dual-stage SEM-ANN analysis," Informatics in Medicine Unlocked, vol. 28, pp. 1-12, 2022.

[14] I. Izonina, R. Tkachenko, "An approach towards the response surface linearization via ANN-based cascade scheme for regression modeling in healthcare," Procedia Computer Science, vol. 198, pp.724–729, 2022.

[15] S. Ellahham , MD, "Artificial intelligence: the future for diabetes care," The American Journal of Medicine, vol. 133, No 8, pp. 895-899, August 2020.

[16] S. Gujral, "Early diabetes detection using machine learning: a review," IJIRST –International Journal for Innovative Research in Science & Technology, vol. 3, Issue 10, pp.57-62, March 2017.

[17] Y.lei SUN, Da-lin ZHANG, "Machine learning techniques for screening and diagnosis of diabetes: a survey," Technical Gazette, vol. 26, no.3, pp. 872-880, 2019.

[18] B. Manoj Kumar P, SrinivasaPerumal R, Nadesh R K, and Arivuselvan K, "Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier," International Journal of Cognitive Computing in Engineering, vol.1, pp. 55-61, 2022.

[19] B. Farajollahi et al., "Diabetes diagnosis using machine learning," Frontiers in Health Informatics, vol. 10, no. 65, pp. 1-5, Mar. 2021.

[20] Henock M. Deberneh and Intaek Kim, "Prediction of type 2 diabetes based on machine learning algorithm," International Journal of Environmental Research and Public Health, vol. 18, pp.1-14, 2021.

[21] Bavkar V C, Shinde A A, "Machine learning algorithms for diabetes prediction and neural network method for blood glucose measurement," INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY, vol. 14, no.10, pp. 869-880, 2021.

[22] V. B. Kolachalama, "Machine learning and pre-medical education," Artificial Intelligence In Medicine, vol. 129, pp.1-3 , 2022.

[23] N. Prerna Tiggaa and Shruti Garga," Prediction of Classification Methods," Procedia Computer Science, vol. 167, pp.706–716, 2020.

[24] R. Budi Lukmanto, Suharjito, Ariadi Nugroho, Habibullah Akbar, "Early detection of diabetes mellitus using feature selection and fuzzy support vector machine," 4th International Conference on Computer Science and Computational Intelligence 2019 (Procedia Computer Science), vol. 157, pp.46–54, 2019.

[25] A. Viloria, Yaneth Herazo-Beltran, Danelys Cabrera, and Omar Bonerge Pineda, "Diabetes diagnostic prediction using vector support machines," The 11th International Conference on Ambient Systems, Networks and Technologies (ANT), April 6-9, 2020, Warsaw, Poland (Procedia Computer Science), vol. 170, pp. 376–381, 2020.

[26] S. Naganandhini and P. Shanmugavadivu, "Effective diagnosis of Alzheimer's disease using modified decision tree classifier," INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, Procedia Computer Science, vol. 165, pp.548–555.

[27] J.Benadit.P and Sagayaraj Francis.F, "Improving the performance of a proxy cache using very fast decision tree classifier," International Conference on Intelligent Computing, Communication & Convergence(ICCC-2015) Bhubaneswar, Odisha, India, Procedia Computer Science, vol. 48, pp. 304-312.

[28] N. Sneha and Tarun Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," Journal of Big Data, pp.1-19, 2019.

[29] P. C. Marques F., "Confidence intervals for the random forest generalization error," Pattern Recognition Letters, vol. 158, pp. 171-175, 2022.

[30] N. Dahiya , B. Saini, and H.D. Chalak, "Gradient boosting-based regression modelling for estimating the time period of the irregular precast concrete structural system with cross bracing," Journal of King Saud University – Engineering Sciences, in press.

[31] John Bonestroo et al., "Forecasting chronic mastitis using automatic milking system sensor data and gradient-boosting classifiers," Computers and Electronics in Agriculture, vol. 198, pp.1-9 , 2022.

[32] A. Yudhana, D. Sulistyo,and Ilham Mufandi, "GIS-based and Naïve Bayes for nitrogen soil mapping in Lendah, Indonesia," Sensing and Bio-Sensing Research, vol. 33, pp.1-11, 2021.

[33] N. Jalal, A. Mehmood, Gyu Sang Choi, Imran Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees," Journal of King Saud University Computer and Information Sciences, vol. 34, Issue 6, Part A, pp. 2733-2742, June 2022.

[34] J. Jamal Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, pp.432–439, 2021.

[35] N. Ahmed et al., "Machine learning based diabetes prediction and development of smart web application," International Journal of Cognitive Computing in Engineering, vol.2, pp. 229–241, 2021.

[36] D. Sisodia and D. Singh Sisodia, "Prediction of diabetes using classification algorithms," International Conference on Computational Intelligence and Data Science (ICCIDS 2018) Procedia Computer Science, vol. 132, pp.1578–1585, 2018.

[37] T. Beghriche , M. Djerioui , Y. Brik , B. Attallah , and S. B. Belhaouari, "An efficient prediction system for diabetes disease based on deep neural network," Complexity, vol. 2021,no.1, pp.1-14, 2021.

[38] Z.C. Sheng , "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," Informatics in Medicine Unlocked. Science Direct. 2019.

[39] S.S. Mirzajani , S. Salimi , "Prediction and Diagnosis of Diabetes by Using Data Mining Techniques," Avicenna Journal of Medical Biochemistry, vol.6(1), pp.3–7, 2018.

[40] M.A.Sarwar , N. Kamal , Hamid W, and Shah MA, Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," In: and others, editor. IEEE 24th International Conference on Automation and Computing. 2018.

[41] A. Singh, M. N. Halgamuge, and R. Lakshmiganthan, "Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 8, no. 12, pp.1-11, 2017