

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/378267434>

Performance Analysis of Deep Neural Network and Machine Learning Algorithms for Diabetes Prediction

Article · February 2024

DOI: 10.1109/ICAHC59020.2023.10431434

CITATIONS

0

READS

117

6 authors, including:



Nrusingha Tripathy

Siksha O Anusandhan University

17 PUBLICATIONS 25 CITATIONS

SEE PROFILE



Gobinda Chandra Das

K L University

5 PUBLICATIONS 5 CITATIONS

SEE PROFILE



Subrat Kumar Nayak

Siksha O Anusandhan University

11 PUBLICATIONS 21 CITATIONS

SEE PROFILE



Sashikanta Prusty

Siksha O Anusandhan University

27 PUBLICATIONS 123 CITATIONS

SEE PROFILE

Performance Analysis of Deep Neural Network and Machine Learning Algorithms for Diabetes Prediction

Nrusingha Tripathy*

Department of Computer Science and Engineering,
Siksha 'O' Anusandhan,
Deemed to be University,
Bhubaneswar, Odisha, India
nrusinghatripathy654@gmail.com

Sarbeswara Hota

Department of Computer Application,
Siksha 'O' Anusandhan,
Deemed to be University,
Bhubaneswar, Odisha, India
sarbeswarahota@soa.ac.in

Pranati Satapathy

Department of IMCA,
Utkal University,
Bhubaneswar, Odisha, India
satapathy.pranati@gmail.com

Gobinda Chandra Das

Department of Computer Science and Applications,
KL Deemed to be University,
Vaddeswaram, Andhrapradesh, India
govindachandrad@gmail.com

Subrat Kumar Nayak

Department of Computer Science and Engineering,
Siksha 'O' Anusandhan,
Deemed to be University,
Bhubaneswar, Odisha, India
subratsilicon28@gmail.com

Sashikanta Prusty

Department of Computer Science and Engineering,
Siksha 'O' Anusandhan,
Deemed to be University,
Bhubaneswar, Odisha, India
sashi.prusty79@gmail.com

Abstract— Diabetes mellitus has an issue called chronic hyperglycaemia. It might lead to a lot of problems. According to present rates of morbidity, it is expected to be 642 million diabetic patients worldwide by 2040, or one in every ten individuals. Without a doubt, more attention has to be paid to this alarming statistic. The huge and highly classified data that the healthcare industry produces must be handled with care. One of the numerous fatal diseases that are spreading around the globe is diabetes mellitus. Medical professionals want a reliable diabetes prediction system. Several machine learning techniques are used in a range of areas to do predictive modelling over huge data. Machine learning is presently applied in many fields of medical science due to its rapid progress. Although it is challenging to employ analytics to predict outcomes in healthcare, in the long run, it may assist practitioners in making quick judgements on the treatment and well-being of individuals based on vast volumes of data. Basically, Pima Indians diabetes dataset is considered in this work. The National Institute of Diabetes and Digestive and Kidney Diseases is the original repository for this dataset. In this work, seven distinct machine learning algorithms are employed and compared with deep neural network model to address the diabetes prediction. The deep neural net model gives better accuracy score as compared to conventional machine learning models.

Keywords—artificial neural network, machine learning, precision, recall, deep learning, diabetes dataset, prediction

I. INTRODUCTION

Diabetes is a communal chronic illness that poses a major threat to people's health. Diabetes is characterised by elevated blood glucose levels, which can be brought on by problems with insulin manufacturing, physiologic effects of insulin, or both. Several tissues, particularly the cardiovascular system, kidneys, liver, nerves, blood vessels, and eyes can suffer persistent damage and malfunction as a result of diabetes. Type 1 diabetes (T1D) and type 2 diabetes (T2D) are the two distinct kinds of diabetes. Type 1 diabetics typically have a younger age than 30 [1], [2]. Among the typical clinical indications are high blood sugar levels, rapid thirst, and frequent urination. Since oral medications alone are unable to cure this kind of diabetes, patients must have insulin therapy.

Type 2 diabetes, which is frequently associated to the occurrence of overweight, hypertension, lipid problems, arteriosclerosis, and other illnesses, is more likely to occur in middle-aged and older adults [3].

Diabetes is growing more prevalent in daily life for individuals as living standards rise. Therefore, it is imperative to do study regarding the simplest and most precise methods of identifying and evaluating diabetes. Resting blood glucose, tolerance to glucose, and variable blood glucose levels are all used in medical diagnosis of diabetes. The easier it is to manage, quicker the diagnosis is determined. Based on data from normal physical tests, users and experts can utilise machine learning to establish an initial diagnosis of diabetes mellitus. The two most significant issues with machine learning is selecting the appropriate classifier and valid features [4]. As one of the most often used machine learning techniques, classification employs it to generate an inferred function that can be used to map new or unseen instances. The present investigation uses classification to create a more accurate prediction model. In order to predict diabetes, predictive machine learning approaches are widely used, and the results are better. A decision tree is a well-liked machine learning technique in the medical industry since it is good at classifying data. Decision trees are produced in great quantity by random forest. The machine learning techniques that have lately gained popularity but the neural network performs better in many ways. So, in this study, we used a variety of classification techniques, regression methods, and neural networks to predict diabetes [5].

II. MATERIALS AND METHOD

The proposed methodology consists of a few steps that are crucial to achieving our objective, such as collecting a diabetes dataset with the relevant patient characteristics, preliminary processing the numerical value of the attributes, using a variety of machine learning techniques, and using the outcome of predictive analysis [6]. The stages are briefly covered in the sections. The overall work flow diagram is given in the Figure 1, where we initially taken diabetic dataset then the data pre-processing occurs, after pre-processing we use correlation analysis of features. Then the different

machine learning models are used for classification and then model evaluation is performed with test dataset. Model prediction occurs on the basis of model accuracy [7].

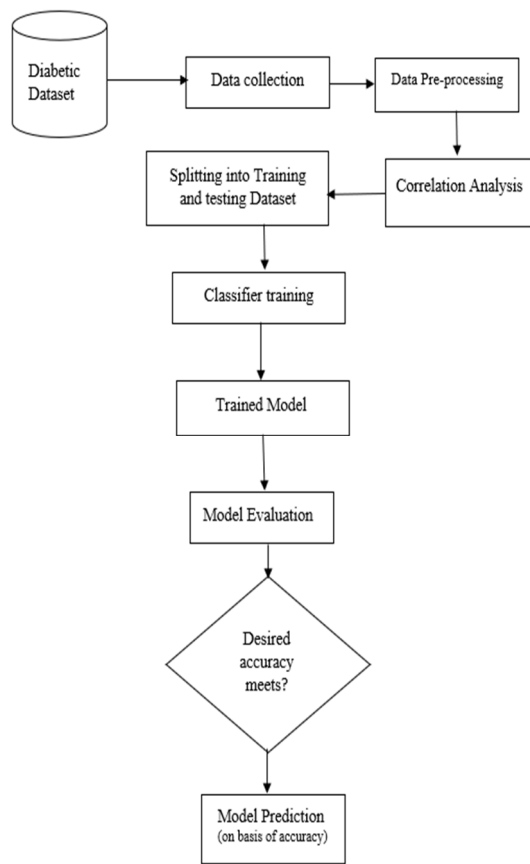


Figure 1. Process Flow Diagram

A. Dataset and Attributes

The National Institute of Diabetes and Digestive and Kidney Diseases served as the original repository for the dataset used in this study. A few diagnostic factors that are part of the collection will be used to figure out a patient's chance of having diabetes. There were a few limitations on the instances that were preferred from a larger database. Every patient at this hospital is a female Pima Indian, in particular.

Table I. Dataset information

Column No.	Column Name	Data type
1	Pregnancies	int64
2	Glucose	int64
3	Blood Pressure	int64
4	Skin Thickness	int64
5	Insulin	int64
6	BMI	float64
7	Diabetes Pedigree Function	float64
8	Age	int64
9	Outcome	int64

These instances were picked under a number of constraints from a larger database. In instance, the bulk of those receiving treatment at this clinic are Pima Indian women under the age of 21. The dataset has 768 instances with a total of 9 characteristics. The datasets consist of many medical predictive (independent) factors and one target (dependent) variable, called Outcome. Independent variables include the patient's BMI, glucose level, age, number of prior pregnancies, and other things. Information about the test dataset and datatype is included in Table I,

B. Data Pre-processing

The diabetes dataset has undergone some data pre-processing in order to meet the objectives of this study. For example, it is useless to forecast diabetes based on the precise numerical value of the features. As a result, we translate into the nominal values of the numerical characteristic. The patient's age, for instance, is alienated into three groups: minor (10–25 years), adult (26–50 years), and old (above 50 years). Similar classifications are made for patients' weights: underweight (less than or equal to 40 kg), normal (41–60 kg), and overweight (more than 60 kg). Last but not least, blood pressure is divided into three groups: Low (less than 80 mmHg), High (more than 120 mmHg), and Normal (120/80 mmHg).

C. Machine Learning Techniques

After the data is ready for modelling, we apply different recognized machine learning classification techniques to forecast diabetes mellitus. As a result of this, we give a summary of alternative approaches.

1. Logistic regression

Calculating the likelihood that the particular instance fits to a specific class is the main use of the automated machine learning technique logistic regression in classification assignments. Classification approaches make use of a technique known as logistic regression. Regression is employed because it uses the output of a linear regression function as input for determining the probability for the given class [8]. In contrast to linear regression, which produces a continuous number that might be anything, logistic regression predicts whether a given instance will belong to a certain class or not.

2. K Nearest Neighbor

Simple regression and classification technique K-nearest neighbour uses non-parametric approach. The algorithm keeps track of all acceptable attributes and categorises novel characteristics based on how similar they are to existing attributes. It employs a tree-like data structure to calculate the distance between the place of interest and the training data set's points. The characteristic is categorised by its surroundings. The value of k in a classification method remains a positive integer of nearest neighbour. An assortment of class or object attribute values are used to select the closest neighbours [9].

3. Classification and Regression Tree (CART)

The CART machine learning technique predicts the values of the target variable based on other characteristics. The needed variable's projected value is found at the final point of every node. The hierarchy of choices is divided into components

that predict at each fork. In the decision tree, sub-nodes are created based on the threshold value of an attribute. The most advantageous parameter and the threshold value are used to split the training set, which serves as the root node, into two groups. The subgroups are also distinguished using the same logic. This process is repeated until the tree produces all of the leaves that are accessible to it or locates the final pure subset [10].

4. Random Forest

A versatile classification and regression method that was created by Dr. Breiman is the random forest algorithm. When there are much more variables involved than different types of observations, the technique has shown to perform well in certain scenarios. It averages the forecasts of several different randomised decision trees. It is an algorithm based on the theory of statistical learning that uses the Bootstrap randomised resampling technique to extract numerous sample sets from the initial training datasets [11]. The algorithm then integrates all of the outputs from the choice trees to forecast the classification outcomes using the pre-established voting mechanism after creating the decision tree framework for each sample set.

5. Support Vector Machine

Some of the most used categorization methods is this one. An occludent classifier known as a Support Vector Machine (SVM) explicitly characterises the data by isolating a hyperplane. SVM separates things within a given class. Instances that are not supported by data can also be recognised and classified [12]. The order of presentation of acquiring data for each class is unimportant to SVM. Using regression analysis to produce a linear function and learning to rank things to provide classification for every component are two ways that this strategy might be improved.

6. XG Boost

Machine learning models may be trained quickly and flexibly using the generalised gradient boosting toolbox known as XGBoost. Using the ensemble learning approach, a number of tentative model predictions are integrated to produce a more solid forecast. One of XGBoost's important characteristics is its effective supervision of values that are missing, which allows it to handle data from the real world with values that are missing with no the need for time-consuming pre-processing. Additionally, XGBoost offers capabilities for parallel processing that enable rapid model training on big data sets. Forecasting of click-through rates, systems for recommendation, and Kaggle contests are just a few uses for XGBoost. Additionally, it is quite adaptable and makes speed optimisation easier because it allows for the fine-tuning of a number of model parameters [13].

7. Light GBM

An efficient version of the widely used Gradient Boosting Decision Tree (GBDT) technique includes XGBoost and parallel Gradient Boosted Regression Trees (pGBRT). Even though both solutions make extensive use of engineering optimisations, their efficiency and scalability are comparatively poor for feature spaces with high dimensions and huge data sets. One key factor is the extremely long calculation time required to test all the data records for each

characteristic in order to assess the information gain of all potential split sites [14]. A gradient boosting system called Light GBM makes use of tree-based learning techniques. Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) are two unique strategies that are used in its distribution and efficiency design. GOSS only utilises the remaining data instances to estimate the information gain after excluding a sizeable part of those with tiny gradients. Since the information gain computation heavily relies on the data records with greater gradients, GOSS can estimate the information gain rather well with a considerably smaller dataset. To decrease the number of features, EFB is used to bundle characteristics that are mutually incompatible [15], [16].

d. Deep Neural Network (DNN)

Deep Neural Network (DNN) are thought to be a mathematically mirror the learning and generalisation capabilities of human neurons. A very nonlinear system with uncertain or complicated variable relationships can be scaled up using an DNN model.

A neural network is made up of layers and different types of neurons. Nodes represent the dendrite and axon portions of the human neuron anatomy, and the connections between nodes represent the weighted axon. The neural network's overall structure is influenced by the data input layer, one or more layers that are concealed, and the output layer. Here, i^{th} neuron represents a connection with j^{th} neuron of the whole structure, and W_{ij} denotes the strength of the link between neurons [17]. The nodes of an DNN's structure receive inputs (features), process them, and then transport the processed data to the next hidden layer through some form of weighted connection, where i^{th} nodes send data to j^{th} nodes for processing, which includes computing the weighting sum and adding up a bias term (θ_j). A mathematical illustration of the ideas above discussed as follows:

$$\text{net}_j = \sum_{i=1}^m x_i * w_{ij} + \theta_j \quad (j = 1, 2, \dots, n) \quad (1)$$

e. Evaluation

This is the prediction model's last phase. Here, we assess the accuracy of the predictions using a variety of measures, such as the classification accuracy, precision, recall, and the F1-score.

1. Classification Accuracy

It measures the proportion number of accurate forecasts to total input trials. It is provided in equation (2),

$$\text{Accuracy} = \frac{\text{Total number of accurate predictions}}{\text{Total number of inputs}} \quad (2)$$

2. Confusion Matrix

A confusion matrix summarises an assessment of a machine learning algorithm's effectiveness on a set of test data. It is commonly employed to evaluate the effectiveness of classification models. Each input event is assigned a category name by these models. The matrix parades how many true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) the model engendered using the test data.

TP	FP
FN	TN

3. Precision

Precision is the proportion of real positive results to all positive results that your model projected as positive results. Precision may be calculated using the formula $TP/(TP+FP)$. You may calculate the rate at which your optimistic forecasts come true using this statistic.

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (3)$$

4. Recall

Recall is a metric that compares the number of real positive outcomes to your genuine positive. Recall may be calculated as follows: $TP/(TP+FN)$. Using this method, we may assess how well our model can identify the real outcome.

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (4)$$

5. F1 Score

A model's recall and accuracy scores are summed up into one metric, the F1 score, which has resulted in its frequent application in recent research [18]. In contrast to accuracy, which measures a model's overall effectiveness, the F1 score measures its capacity to forecast by concentrating on how well it succeeds within each class.

$$F1 = 2 * \frac{1}{\left(\frac{1}{\text{precision}}\right) + \left(\frac{1}{\text{recall}}\right)} \quad (5)$$

III. CORRELATION MATRIX OF DATASET

A metabolic condition called diabetes mellitus causes unusually elevated levels of sugar in the blood. A group of metabolic diseases known as diabetes are typified by chronically increased blood sugar levels. High blood sugar is indicated by frequent urination, increased thirst, and increased hunger. If diabetes is neglected, a variety of negative effects may result. Acute consequences comprise hyperosmolar hyperglycaemia, diabetic ketoacidosis, and even death. Serious long-term effects include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, and vision impairment. Insulin is a hormone that helps cells store or use blood sugar for energy [19]. Diabetes is characterised by either insufficient or ineffective insulin production by the body. Your nerves, eyes, kidneys, and other organs may be harmed if high blood sugar from diabetes is left untreated.

According Figure 2 it shows the Correlation matrix graph of the data set. Heatmaps make it extremely simple to understand the association between one characteristic (variable) and every other feature (variable). To put it another way, a correlation matrix is a table of data that shows the 'correlations' between sets of variables in a set of data.

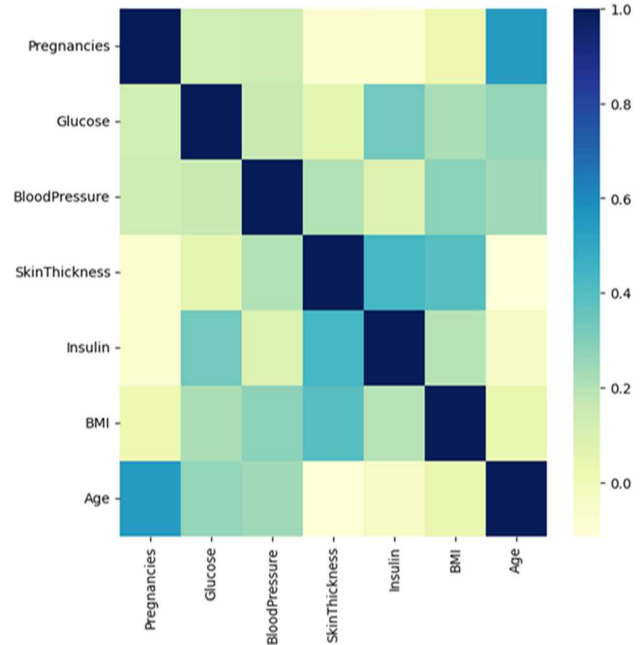


Figure 2. Correlation matrix graph of the data set

There is a battery of examinations and tests that must be finished in order to correctly diagnose diabetes in persons who may have it. These evaluations could include redundant or unneeded medical procedures, which creates issues and wastes time and resources. Diabetes has a considerably greater economic impact than it does through medical expenditures to the healthcare system since it decreases quality of life and reduces productivity at work. Figure 3 shows the visualization of missing observations, it happens frequently when using a dataset from the actual world that certain values are missing. These missing values are filled with the median values.

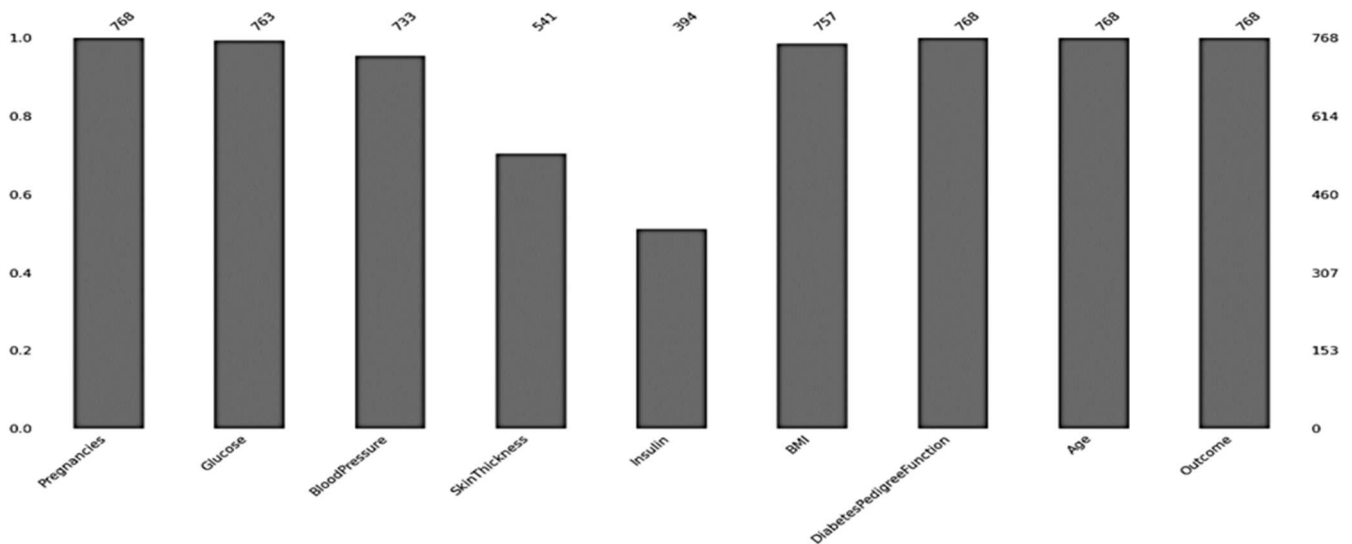


Figure 3. visualization of missing observations

Lack of an appropriate diagnosis plan, a lack of funding, and a general lack of information are the main contributors to these negative effects. Therefore, avoiding the illness completely by early identification may undoubtedly reduce a substantial economic load and assist the patient in managing their diabetes. The suggested system's advantages include, but are not limited to, the understated,

- Regarding the hospital management, it would serve as a Decision Support System (DSS), which would greatly aid them in making a quick choice that is of high quality.
- By using this approach, it is hoped that more effective measures may be implemented to lessen the negative effects that diabetes has on the patient and to provide recommendations that would enable the patient to properly manage his health.
- Since most operations would be automated under the proposed system, it saves the hospital administration the time and effort used to create patients' diseases in the current system.

IV. RESULT ANALYSIS

Seven machine learning algorithms were employed in this experiment and is compared with DNN model. The machine learning models are LR, KNN, CART, RF, SVM, XGB and LightGBM algorithms. These methods were all used using the PIMA Indian diabetes dataset. Training data and testing data were divided into two groups, each of which contained 70% and 30% of the total data. Prediction accuracy was our main criterion for evaluation in this study. The algorithm's overall success rate is known as accuracy. Algorithm accuracy with Precision, Recall and F1 Score was evaluated and shown in Table II. The accuracy and parameter values given in the Table, are plotted through bar diagram for easy visualisation are given in the Figure 4. The box plot of the accuracy values of the classifiers is also shown in Figure 5. The results indicates that the accuracy of DNN is 0.92 which is better than all other models. This DNN model has one input layer, seven hidden layers and one output layer. The relu and sigmoid activation functions have been used with the loss function as binary cross entropy.

Table II. Model evaluation on the basis of accuracy and different parameters

Models	Accuracy score	Precision	Recall	F1 Score
LR	0.84	0.78	0.93	0.86
KNN	0.84	0.78	0.93	0.86
CART	0.85	0.86	0.94	0.86
RF	0.88	0.88	0.94	0.89
SVM	0.85	0.86	0.94	0.86
XGB	0.89	0.89	0.94	0.89
LightGBM	0.88	0.88	0.94	0.89
DNN	0.92	0.89	0.95	0.92

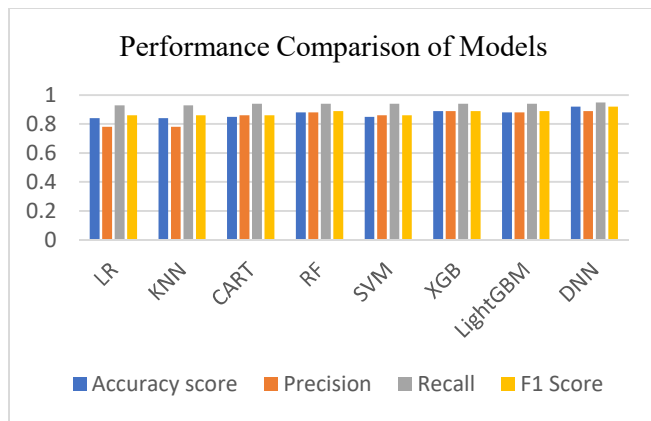


Figure 4. Model comparison on the basis of different parameters

V. CONCLUSION

The dataset in this study is subjected to a diversity of machine learning techniques, and numerous algorithms have been used to classify the data. As compared to traditional machine learning models the Deep Neural Network gives better result and the result is obvious because the database is not used in traditional programming; instead, data is stored over the whole network. Although some data briefly vanishes from one point, the network still works. There are several possible negative outcomes associated with diabetes mellitus. It could be beneficial to look into applying machine learning to accurately predict and diagnose this illness. It implies that diabetes may be predicted using machine learning, but it's critical to choose the appropriate features, classifier, and data mining approach. We want to determine the kinds of diabetes in the future and also analyse the relative significance for every signal, which may enhance the precision with which diabetes is forecast because we cannot identify the kind of diabetes based just on the data. Additionally, it is possible for non-diabetic persons to develop diabetes in the coming years.

Reference

- [1] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong and G.-Z. Yang, "Big Data for Health", *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 4, pp. 1193-1208, 2015.
- [2] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", *IEEE Access*, vol. 5, no. c, pp. 8869-8879, 2017.
- [3] J. B. Heaton, N. G. Polson and J. H. Witte, "Deep learning for finance: deep portfolios", *Appl. Stoch. Model. Bus. Ind.*, vol. 33, no. 1, pp. 3-12.
- [4] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data from Healthcare Communities", *IEEE Access*, vol. 5, pp. 8869-8879, 2017.
- [5] N. Tripathy, S. Hota, S. Prusty, and S.K.Nayak. "Performance Analysis of Deep Learning Techniques for Time Series Forecasting." In *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*, pp. 639-644. IEEE, 2023.
- [6] D. M. Renuka and J. M. Shyla, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", *Int. J. Appl. Eng. Res. ISSN*, vol. 11, no. 1, pp. 973-4562, 2016.
- [7] K. Kayaer and T. Yildirim, "Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks", *International Conf Artif. Neural Networks Neural Inf. Process.*, pp. 181-184, 2003.
- [8] M. A. Abdul-Ghani and R. A. DeFronzo, "Plasma Glucose Concentration and Prediction of Future Risk of Type 2 Diabetes", *Diabetes Care*, vol. 32, no. suppl_2, pp. S194-S198, Nov. 2009.
- [9] Z. Punthakee, R. Goldenberg and P. Katz, "Definition Classification and Diagnosis of Diabetes Prediabetes and Metabolic Syndrome", *Can. J. Diabetes*, vol. 42, pp. S10-S15, 2018.
- [10] G. Swapna, R. Vinayakumar and K. P. Soman, "Diabetes detection using deep learning algorithms", *ICT Express*, vol. 4, no. 4, pp. 243-246, 2018.
- [11] S. K. Nayak, A. K. Nayak, S. Mishra, P. Mohanty, N. Tripathy, A. Pati and A. Panigrahi, "Original Research Article Speech data collection system for KUI, a Low resourced tribal. Journal of Autonomous Intelligence", 7(1).
- [12] M. T. P. Kamble, "Diabetes Detection using Deep Learning Approach", vol. 2, no. 12, pp. 342-349, 2016.
- [13] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, "Predicting Diabetes Mellitus with Machine Learning Techniques", *Front. Genet*, vol. 9, pp. 1-10, 2018.
- [14] N. Tripathy, S. Hota and D. Mishra, "Performance analysis of bitcoin forecasting using deep learning techniques", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 31, no. 3, pp. 1515-1522, 2023.
- [15] V. A. K. and R. C., "Classification of Diabetes Disease Using Support Vector Machine", *International Journal of Engineering Research and Applications*, vol. 3, pp. 1797-1801, April 2013.
- [16] N. Tripathy, S. Parida and S. K. Nayak, "Forecasting Stock Market Indices Using Gated Recurrent Unit (GRU) Based Ensemble Models: LSTM-GRU", *International Journal of Computer and Communication Technology*, vol. 9(1), 2023.
- [17] X. H. Meng, Y. X. Huang, D. P. Rao, Q. Zhang and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors", *The Kaohsiung journal of medical sciences*, vol. 29, no. 2, pp. 93-99, 2013.
- [18] A. Swarupa Rani and S. Jyothi, "Performance analysis of classification algorithms under different datasets", *Computing for Sustainable Global Development (INDIACom) 2016 3 rd International Conference on*, pp. 1584-1589, 2016.
- [19] N. Tripathy, S. K. Nayak, J. F. Godslove, I. K. Friday, and S. S. Dalai, "Credit Card Fraud Detection Using Logistic Regression and Synthetic Minority Oversampling Technique (SMOTE) Approach", *International Journal of Computer and Communication Technology*, 8(4), p.4, 2022.

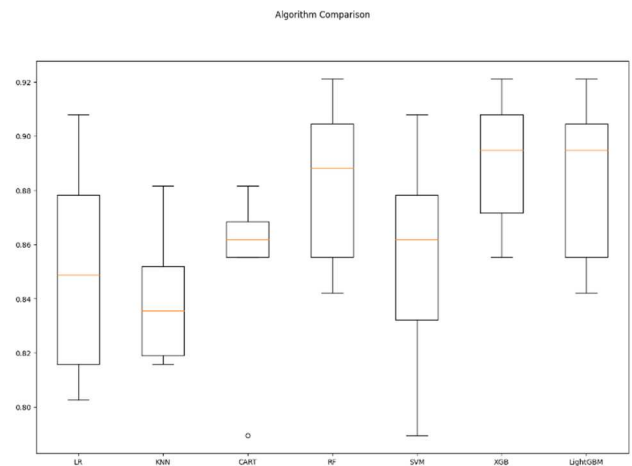


Figure 5. Box plot of the machine learning models