http://www.jctjournals.com ISSN(ONLINE):2278 – 3814

Volume-15, Issue-05, May 2026
JOURNAL OF COMPUTING TECHNOLOGIES (JCT)
International Journal
Page Number: 01-06

# Healthcare Efficient Diabetes Disease Prediction Using Supervised Machine Learning

**Saurabh Kumar Pandey[1], Prof. Preeti Mishra [2]**
[1]M.Tech Scholar, [2]Assistant Professor,
[1,2] Mittal Institute of Technology, Bhopal, Bhopal, INDIA
Email ID: Saurabhkumarpandey70913@gmail.com, pritypreet85@gmail.com

***Abstract*— One of the major threats to human health today is Diabetes Mellitus (DM). Diabetes is a metabolic disease where a person suffers from increased sugar levels because either the pancreas does not produce sufficient insulin for the body or the cells do not respond to the insulin. Persistent Diabetes leads to malfunction, injury and failure of organs such as kidneys, eyes, nerves, blood vessels and heart. The clinical data includes several tests needed to diagnose Diabetes mellitus, depending upon the healthcare personal experience. Hence, it is essential to predict the symptoms of Diabetes at its beginning stage to prevent its growth with appropriate medical diagnosis. These symptoms need to be wisely classified for early detection of Diabetes. Hence, the design of a classifier is essential for detecting Diabetes disease with optimal cost and better performance. In this research, the main objective is to classify the data as diabetic or non-Diabetic and improve the classification accuracy. It presents an automatic prediction system for Diabetes mellitus through machine learning techniques. It considers several limitations of traditional classifiers and provides a significant relationship between patients' symptoms with diabetes diseases and the blood sugar rate. Machine learning offers reliable and excellent support for predicting a DM with the correct case of training and testing. Diagnosis of Diabetes mellitus requires good support of machine learning classifiers to detect diabetes disease in its early stage since it cannot be cured at later stages and subsequently bring more complications to a person's health system. In this thesis, the Gradient boosting machine learning technique is implemented to train the Diagnosis of Diabetes and to classify the diabetes patients in two class values. Positive diabetes patients are defined by class '0' value, and negative diabetes patients are defined by class' 1'. The total Diagnosis diabetes dataset is 1145. All datasets applied to the Gradient boosting machine learning technique are divided into two groups. The first group consists of 815 datasets belonging to non-Diabetes, and the second group includes 330 datasets of Diabetes.**

***Keywords*— *Machine Learning, Supervised Learning, Diabetes Mellitus, Health System, Gradient Boosting, Python, Colab Platform, Recall, Accuracy.***

## I. INTRODUCTION

Diabetes mellitus, characterized by elevated blood glucose levels (hyperglycemia), is a chronic condition posing significant global health challenges. Projections indicate a concerning trend, with an estimated 642 million individuals worldwide expected to be affected by diabetes by 2040, translating to approximately one in ten adults.1 The hormone insulin plays a pivotal role in regulating blood glucose levels, and insufficient production or ineffective utilization of insulin are primary factors contributing to diabetes. Left unmanaged, diabetes can lead to a myriad of health complications affecting multiple organs within the body.2 Alarmingly, the prevalence of diabetes is on the rise, with an estimated 452 million cases identified in 2017 and projected to escalate to 694 million by 2045 (Lawrence e.al., 2021). Moreover, studies suggest that by 2030 and 2045, diabetes prevalence is anticipated to encompass 25 % and 51 % of the population, respectively.3

The integration of machine learning techniques has significantly influenced various fields, with medicine being a primary beneficiary4, 5, 6). Leveraging insights derived from extensive data analysis, these technologies facilitate the creation of more precise and personalized approaches to diagnosis and treatment. Consequently, healthcare providers are empowered to make informed decisions promptly, enhancing patient care outcomes.7 For instance, within oncology, machine learning models can predict treatment responses and customize therapeutic strategies according to each patient's unique characteristics, thereby optimizing treatment efficacy while minimizing adverse effects.

Various algorithms are employed for diabetes prediction, encompassing traditional machine learning methods such as Support Vector Machine (SVM), Decision Tree (DT), and Logistic Regression.8 Polat & Günes9 differentiated between diabetes and normal individuals using Principal

 

Component Analysis (PCA) and Neuro-Fuzzy Inference. Recent research on diabetes prediction has seen the widespread adoption of a diverse range of machine learning models, spanning from conventional algorithms like Logistic Regression and k-Nearest Neighbors to advanced techniques such as Artificial Neural Networks, Random Forests, and Deep Neural Networks. Darolia & Chhillar10 examined a diabetes dataset utilizing popular algorithms like Artificial Neural Network, Random Forest, and Logistic Regression. Notably, their findings demonstrated that Logistic Regression outperformed other algorithms, underscoring its effectiveness in diabetes prediction. Febrian et al.11 took a different approach, employing supervised machine learning techniques and comparing two k-Nearest Neighbor algorithms against the Naive Bayes algorithm for diabetes prediction. Interestingly, their study concluded that the Naive Bayes algorithm exhibited superior performance compared to KNN in this context. Mousa et al.12 conducted a comprehensive comparative study focusing on three widely used models: Long Short-Term Memory, Random Forest, and Convolutional Neural Network for diabetes diagnosis. The article is structured as follows: Section 2 reviews recent literature on the use of algorithms for modeling various healthcare data. Section 3 explains the performance measures and the machine learning algorithms used to model the diabetic data. Section 4 provides an analysis of the compared algorithms. Finally, Section 5 presents the conclusions and recommendations.

## II. LITERATURE REVIEW

Recent advancement in technology created a world of data that ultimately enables both practical and experimental application of machine learning in health care. Machine learning has been observed to be very useful in diagnosis and prediction of different health challenges whether infectious or non-infectious.13 Machine learning algorithms make use of AI concepts which are generally classified as supervised, unsupervised and reinforcement learning. The supervised learning approach make use of data to retrieve information from the training set and conceptualize models that can accurately predict the training set outcomes and uses the train model to make predictions of new features in the testing data set.14 Logistic regression, Decision Trees, Support Vector Machines, Artificial Neural Networks and Random Forest Regression are some of the examples of machine learning algorithms that are based on supervised learning approaches.15 In an unsupervised learning approach, the algorithms are used to group data into independent clusters which makes it easy for features extraction and classification. In contrast to other forms of approaches in machine learning, unsupervised learning allows for extraction of features by identifying the relationships within the data points and grouping them into clusters based on their similarities. Common examples of unsupervised learning include K-Means, KNN, Deep Belief Network (DBN) and CNN.13 Reinforcement learning techniques are the most efficient form of machine learning algorithms that is closest to human and animal intelligence.17 This approach works based on self-learning to eliminate error and improve its overall model performance. The most common reinforcement learning method is the Recurrent Neural Network. Scientists and researchers make use of many machine-learning approaches in diagnosis and prediction of several health challenges to come up with an innovative and better performance solutions compared to the traditional ways of disease diagnosis and treatment in healthcare. They apply the Machine learning algorithms to Electronic Health Records (EHR), Medical Imaging, and genetic engineering to solve different health problems, forecast disease spread and make recommendations about prevention and control mechanisms. El-Bashbishy and El-Bakry proposed a novel technique for early diabetes prediction with high accuracy. They optimized data preprocessing, prediction, and classification using a novel dataset of Mansoura University Children's Hospital Diabetes (MUCHD), which allowed for a comprehensive evaluation of the system's performance. Various validation metrics were employed to ensure the reliability of the results using cross-validation approaches with various statistical measures of accuracy, F-score, precision, sensitivity, specificity, and Dice similarity coefficient. introduced the first-ever self-explanatory interface for diagnosing diabetes patients using machine learning. They proposed four classification models (Decision Tree (DT), K-nearest Neighbor (KNN), Support Vector Classification (SVC), and Extreme Gradient Boosting (XGB)) based on the publicly available diabetes dataset. All the models exhibited commendable accuracy in diagnosing patients with diabetes, with the XGB model showing a slight edge over the others.

Liu et al. developed a deep learning algorithm model using a combination of reinforcement and supervised learning approaches to accurately predict the beginning of various common diseases such as stroke, kidney failure, and heart failure. The prediction model makes use of both structured and unstructured data from EHR and diagnosis notes which resulted in the model high performance and versatility. In another research using EHR data, Ahmad and Ali21 predict mortality in paralytic ileus (PI) - a medical condition characterized by incomplete blockage of the intestine that prevent direct passage of food substance and which may later lead to total blockage of the intestine. In this research, they come up with an algorithm that predict the mortality in PI patients with an accuracy of 81.3 %. This development leads to improved awareness as well as robust clinical treatment for the ailment.

Medical imaging is another important health field that has and continues to witness the transformative application of machine learning applications in healthcare. Machine learning applications and techniques have been employed to improve imaging modalities in areas like Computed Tomography (CT), Ultrasound, X-Ray, Magnetic Resonance Imaging (MRI) to mention a few.3 With Medical imaging, McKinney et al.22 recently developed a deep learning algorithm to detect tumors in Mammogram images in the early stage of cancer. A careful comparison to assess the performance of this algorithm shows that it outperforms an experienced radiologist who use the

 

traditional method of tumors detection by 11.5 %. In the same vein, Esteva et al.23 make use of a convolutional neural network for a classification of 2032 different skin diseases with dermoscopic images. The performance of the Classification algorithms compares with practical method by 21-board member of registered dermatologist give similar results. Diabetic Retinography is a common eye condition that usually affects up to 60 % of type 1 diabetic patient, which is difficult to detect at the early stage.24 Alqudah25 applied a deep learning CNN to detect the aneurysms that cause impaired eye vision in diabetic retinography progression. The early detection of this abnormality through the CNN approach makes it possible to prevent irreversible damage to the patient's vision.

## III. PROBLEM FORMULATION

In the present era, diabetic disease results in increase within the death rate in most of the country. Diabetes disease is becoming a common disease occurring to humans due to an inadequate production of insulin and also due to high amount of sweetness in the body fluids. Before generating the clinical examination, several symptoms are adopted in cause of diabetic disease.

- To effectively diagnose a health issue, proper care must be taken while considering the relevant parameters such as the patient's daily routine, eating habits, medical history, etc. For efficient use of machine learning in predicting and diagnosing the health problem, the above-mentioned parameters are necessary as an input variable for successfully formulating an algorithm. The problem formulation is carried out as follows:
- A most effective tool is needed for early-stage prediction and diagnosis of disease such as Diabetes Mellitus, Cancer, Heart problems etc., with the highest accuracy and lowest misclassification.
- In medical science, disease data diagnosis involves many medical tests that is needed to diagnose a specific disease. The effective diagnosis depends on the health care personal experience since a less experience person can misdiagnose a health problem.
- The number of false positives is quite high in some specific cases, which can be further reduced using the Machine learning algorithm.
- The Classifier to be designed should be efficient, convenient, and, most importantly, must be capable of classifying and predicting the Diabetes patient with the highest accuracy and minimal misclassification.

## IV. PROPOSED MODEL

The primary goal of this work is to categories data as diabetes or non diabetic and to enhance classification accuracy. It offers an automated diabetes mellitus prediction system based on machine learning methods. With the proper scenario of training and validation, machine learning offers a dependable and great assistance for DM prediction. Diabetes mellitus detection requires significant assistance from machine learning classifiers to identify the illness at a preliminary phase, as it cannot be treated, posing significant complications for our health system. This study aided in the development of a classification system for DM prediction.

### *A. Proposed Methodology*

**XG Boosting Classifier** XG Boosting is an effective machine learning method that may be used for a variety of classification problems, including the classification of breast cancer. With the use of the ensemble learning technique known as gradient boosting, a powerful predictive model is produced by combining the predictions of several weak learners, often decision trees. It's a preferred option because it frequently produces high accuracy and can manage intricate data interactions.

### *B. Flowchart*

Making a classification model out of a dataset with labelled classes and some features, like a dependent binary variable and an independent variable, is the main objective of machine learning techniques. The majority of the XG boosting machine algorithms' workflow is composed of the training and dataset validation phases. Using the training dataset, the method adjusts the prediction model to reduce error in the output results.

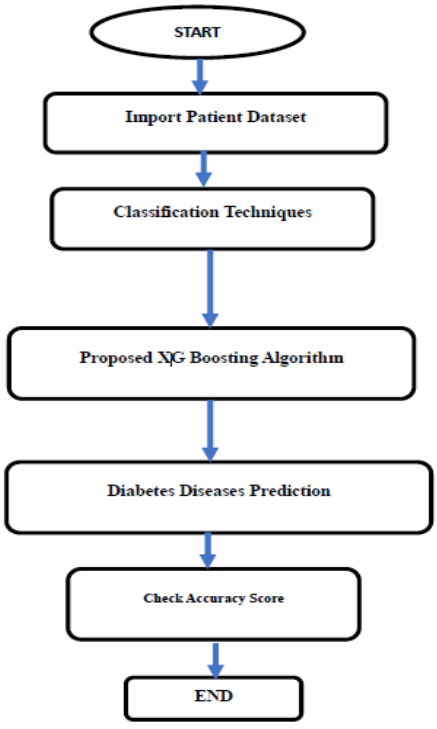


**Fig. 1:** Flow chart of Proposed Algorithm

## V. SIMULATION RESULT ANALYSIS

Python tool with colab platform is used for this project. All of them are free and open source. Python is a general programming language and is broadly utilized in

 

a wide range of disciplines like general programming, web improvement, programming advancement, information investigation.

### *A. Data Set*

The investigation is based on the U.C.I. machine learning store's diabetic Pima Indian dataset, which contains 1145 information occurrences and 9 traits.

### *B. Dataset Description*

The investigation is base1 on the U.C.I. machine learning store's diabetic Pima Indian dataset [12], which contains 1145 information occurrences and 9 traits. All of the women in this dataset are Pima Indians and are at least 21 years old. Their age is indicated by either a "0" or a "1," with a "0" indicating a negative test for diabetes and a "1" indicating a positive test. Table 5.1 shows the number of features, classes and patterns and table 5.2 represent the confusion matrix of prima Indians diabetes dataset.

### *C. Result Analysis*

Machine Learning is a thought that agrees over the machine to take data from instances and former knowledge, and learn from historic data to make predictions based on the learning of the past data and that too without being programmed by any programmer i.e. we can use previous data for future predictions. In this case, instead of programmer writing the code, what a naïve user can do is feeding data to the generic algorithm, and the logic is build based on trained data by the algorithm/ machine. For e.g. When we shop online, while looking for a product, we have noticed that similar products are recommended to us to what we were looking for and we also notice the following quotation "the person who purchased this product also purchased this" type of combination of products. This recommendation is done using machine learning. Many a times we get a phone call from the bank or the finance company asking us to take a loan or purchase an insurance policy.

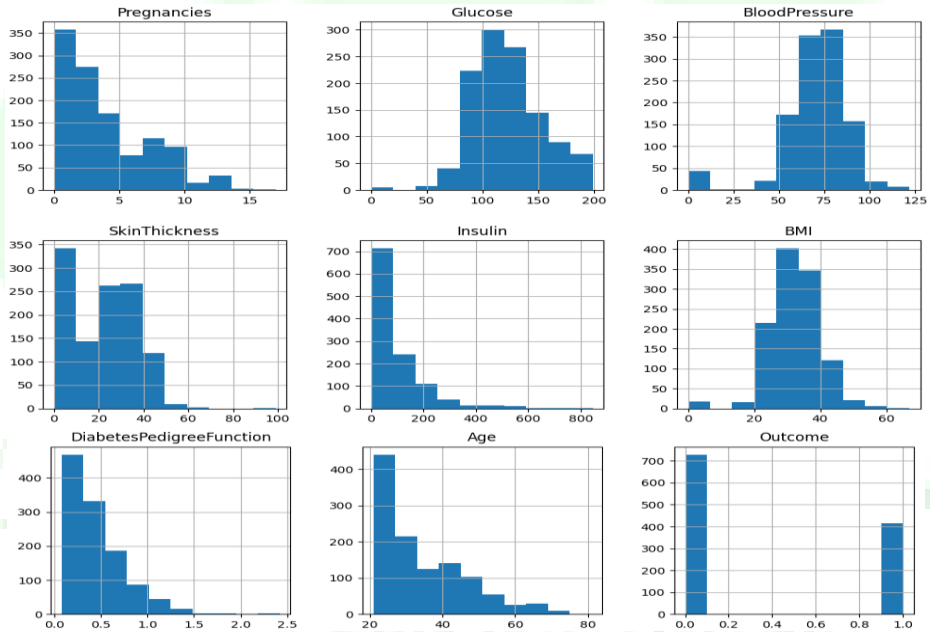


**Fig. 2:** Histogram of Dataset

Fig. 2 and 3 show the status of Diabetes health, ranging from healthy to severely unhealthy. Blue bar represents

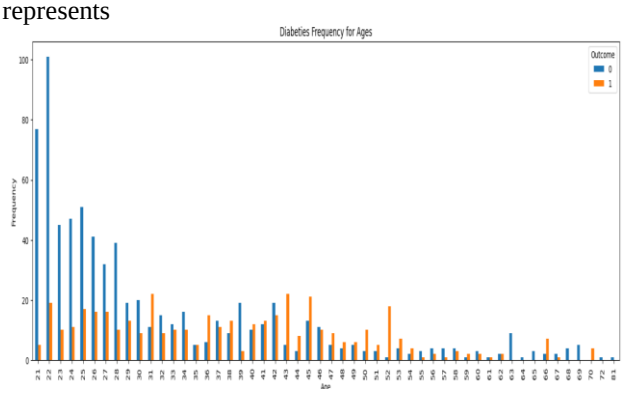


**Fig. 3:** Bar Plot of the Number of Diabetes Frequency for Ages

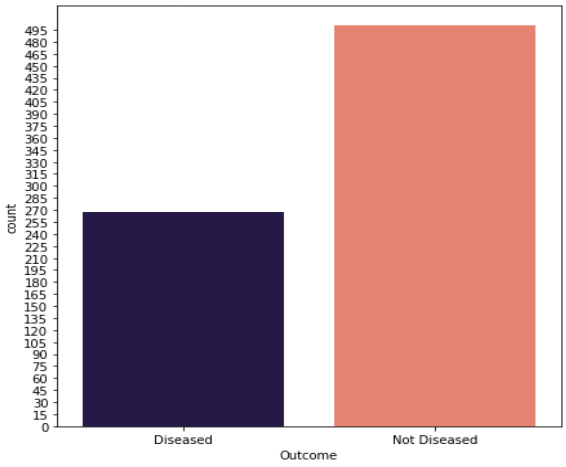


**Fig. 4:** Bar Plot According to Outcomes

 

Machine Learning algorithms are trained by using a training data set for model creation. When some new input data from the attributes is feed to the algorithm of machine learning, prediction is done on the basis of the selected model is shown in Fig. 4. Then the predictions are measured for accuracy. If the accuracy of the input data is acceptable, deployment of the Machine Learning algorithm is done on input data. If the accuracy of input data is not acceptable, the algorithms with some new data are trained again and again with an arbitrary training data set.

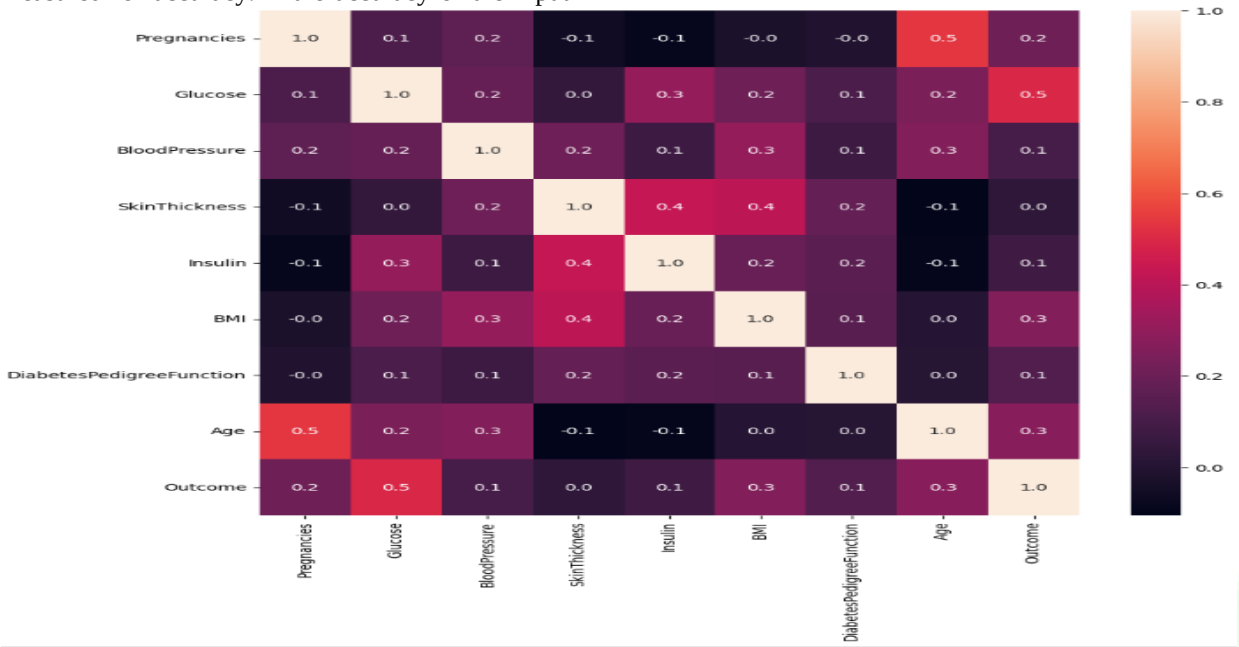


**Fig. 5:** Bar Plot According to Diabetes Pedigree Function

In proposed demonstrating at least two related however unique scientific models are utilized and produce their outcomes are joined into a solitary score.

Subsequent to playing out the AI approach for testing and preparing we find that exactness of the Inclination supporting is much proficient when contrasted with different calculations. Precision ought to be determined fully backed by disarray framework of every calculation as displayed in Figure 5.5, here number of counts of T.P., TN, F.P., F.N. are given and utilizing the condition of precision, esteem has been determined and it is presume that proposed calculation is best among them with 92.18% exactness and the correlation is displayed in Table I.

Table I: Assessment of Different Classification Methods

| Sr. No. | Algorithm | Accuracy |
|---|---|---|
| 1 | LR | 74.82% |
| 2 | GNB | 73.42% |
| 3 | RFC | 83.56% |
| 4 | K-NN | 71.32% |
| 5 | DT | 80.76% |
| 6 | SVM | 82.51% |
| 7 | Proposed Algorithm | 91.23% |

Table 5.5 represents the accuracy for different ML classifier with Previous Isfafuzzaman Tasin et al. [1]. GB classifier is best accuracy compared to Previous Isfafuzzaman Tasin et al. [1].

Table II: Comparison Result for Accuracy

| Techniques | Previous Isfafuzzaman Tasin et al. [1] | Proposed Algorithm |
|---|---|---|
| DT | 72% | 80.76% |
| K-NN | 73% | 71.32% |
| LR | 75% | 72.82% |
| RF | 76% | 83.56% |
| SVM | 78% | 82.51% |
| NB | 79% | 73.42% |
| GB | 81% | 91.23% |

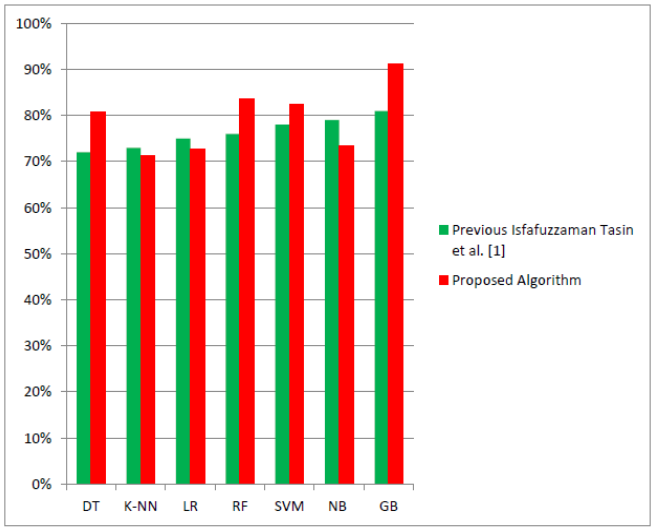


**Fig. 6:** Bar Graph of the Previous and Proposed Algorithm for Accuracy

 

## VI. CONCLUSION AND FUTURE WORK

One significant health challenge worldwide is Diabetes detection early and proactively. The present study aims to establish a suitable prediction model that depends on a Machine Learning scheme for predicting Diabetes. Diabetes is a metabolic condition brought on by elevated levels of glucose in the blood. Patients with diabetes don't have enough insulin in their bodies to control their sugar levels. Additionally, diabetes contributes to a number of other dangerous diseases. Due to the fact that diabetes is the root cause of numerous diseases in the human body, it is essential to detect this serious condition as soon as possible.

There have been a number of approaches taken in the past to make it easier for doctors to diagnose diabetes. However, there is a classification issue with it: a selection of an excessive number of samples, but it does not improve classification accuracy. The speed of the algorithm is better in a few situations, but the accuracy of the data classification is lower. A selected classification algorithm is used to test the diabetic data set. We applied a gradient-boosting machine learning algorithm to the Diabetes dataset's attributes in order to get the best results.

The diabetes dataset of 1145 Pima Indians: The test uses 330 diabetic and 815 non-diabetic participants. An ensemble of gradient boosting was used in the proposed algorithm to achieve an accuracy of 91.23%. As can be seen, the majority vote-based model employs NB, DT, and SVM classifiers, and its accuracy for the diabetes disease dataset is 73.42%, 80.76%, and 82.51%, respectively. Subsequently, the Inclination helping calculation gives the best exactness to Diabetes finding than the past calculation.

We propose two potential headings for future examination.

- Only supervised learning methods and predefined health datasets are used in the proposed approach. This structure will deal with electronic wellbeing records utilizing solo learning approaches in future.
- The proposed work may be expanded to include unstructured and semi-structured information in the future, but it currently only addresses structured data.

**References**


[1]. Tansin Ullah Nabil, Sanjida Islam, Riasat Khan, “Diabetes prediction using machine learning and explainable AI techniques”, Healthcare Technology Letters, pp. 01-10, Wiley 2022.

[2]. Olisah, C.C., Smith, L., Smith, M., “Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective”, Comput. Methods Programs Biomed., Vol. 20, pp. 1–12, 2022.

[3]. Deberneh, H.M., Kim, I., “Prediction of type 2 diabetes based on machine learning algorithm”, Int. J. Environ. Res. Public Health, Vol. 18, pp. 1–14, 2021.

[4]. Nikos Fazakis, Otilia Kocsis, Elias Dritsas, Sotiris Alexiou, Nikos Fakotakis, and Konstantinos Moustakas, “Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction”, IEEE Access 2021.

[5]. Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy, “Prediction of Diabetes using Machine Learning Classification Algorithms”, International Journal of Scientific & Technology Research, Vol. 9, No. 01, 2020.

[6]. Chatrati, S.P., Hossain, G., Goyal, A., “Smart home health monitoring system for predicting type 2 diabetes and hypertension”, J. King Saud Univ. Comput. Inf. Sci., Vol. 34, No. 3, pp. 862–870, 2020.

[7]. Hasan, M.K., Alam, M.A., Das, D., Hossain, E., Hasan, M., “Diabetes prediction using ensembling of different machine learning classifiers”, IEEE Access, Vol. 8, pp. 76516–76531, 2020.

[8]. Cervantes, J., García-Lamont, F., Rodríguez, L., Lopez-Chau, A., “A comprehensive survey on support vector machine classification: Applications, challenges and trends”, Neurocomputing, Vol. 408, pp. 189–215, 2020.

[9]. Pranto, B., “Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh”, Information Vol. 11, pp. 1–20, 2020.

[10]. Mohan, N., Jain, V., “Performance analysis of support vector machine in diabetes prediction”, In: International Conference on Electronics, Communication and Aerospace Technology, pp. 1–3, 2020.

[11]. Muhammad Azeem Sarwar, 2Nasir Kamal, 3Wajeeha Hamid,4Munam Ali Shah, “Prediction of Diabetes Using Machine Learning Algorithms in Healthcare”, 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2019.

[12]. Rao G.A., Syamala K., Kishore P.V.V., Sastry A.S.C.S. ., “Deep convolutional neural networks for sign language recognition”, International Journal of Engineering and Technology(UAE) ,Vol: 7, Issue 5, pp: 62-70, 2018.

[13]. Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju and Hua Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques”, Springer, 2018.

[14]. L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, “Machine learning on big data: Opportunities and challenges,” Neuro computing, vol. 237, pp. 350–361, May 2017.

[15]. 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, Jan. 2017.


 