



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.524

Volume 14, Issue 1, January 2025

Prediction of Diabetes Risk and its Effect Health Care System using XG Boosting Technique

Tarun Tapan Singh¹, Prof. Suresh S. Gawande²

M. Tech. Scholar, Department of Electronics and Communication, Bhabha Engineering Research Institute,
Bhopal, India¹

Guide, Department of Electronics and Communication, Bhabha Engineering Research Institute, Bhopal, India²

ABSTRACT: Diabetes is a major global health concern, leading to severe complications and increased healthcare costs if not diagnosed and managed promptly. Early prediction of diabetes risk is critical for implementing preventative measures and optimizing healthcare resources. This study focuses on leveraging the eXtreme Gradient Boosting (XGBoost) technique, a robust supervised machine learning algorithm, to predict diabetes risk with high accuracy and reliability. XGBoost's ability to handle imbalanced datasets, process large volumes of data, and deliver interpretable results makes it an ideal choice for healthcare applications. The proposed model utilizes clinical, demographic, and lifestyle-related data to predict diabetes risk effectively. It emphasizes feature engineering and selection to optimize model performance while evaluating results using metrics such as accuracy, precision, recall, and ROC-AUC. Furthermore, the study explores the broader impact of diabetes risk prediction on the healthcare system, including resource allocation, personalized care, and cost reduction through early intervention. This research aims to provide a comprehensive framework for integrating XGBoost-based predictive models into healthcare systems, contributing to improved decision-making and better health outcomes for individuals at risk of diabetes. The findings highlight the potential of machine learning in transforming preventive healthcare strategies and fostering a more efficient and patient-centric approach.

KEYWORDS: Diabetic dataset, Classification, Machine Learning, Gradient Boosting

I. INTRODUCTION

Diabetes is an extreme growing health problem in all over the world which causes death, especially in industrial and developing countries. This chronic disease may over to long term complications and death. It can cause high risk of kidney failure, nervous system damage, blindness and heart diseases. In this disease the body does not growth or rightly use the insulin, the hormone that release the cells of the body, grant the glucose to get in and ammunition them [1].

In the absence of insulin the cells become starving of glucose energy against the being of abundant glucose in blood stream. Complications of diabetes are linked to blood vessel diseases and generally classify into bitty vessel disease such as those affect the eyes, called as diabetic retinopathy. It occurs in the patients who have diabetes for minimum five years with the drainage of protein in small blood vessels at the back of eyes and the blood in retina. Disease in blood vessel also causes the formation of small aneurysms and new but brittle blood vessel led to retinal scarring and retinal detachments thus impairing vision [2].

Kidney damage from diabetes is called diabetic nephropathy. Initially disease blood vessel in the kidney causes the leakage protein in the urine. Later on kidneys lose their capacity to scrub and channel blood. The collection of dangerous waste item in blood prompts the requirement for dialysis.

Nerve harm from diabetes is called diabetic neuropathy and it is additionally caused by ailment of little veins. The blood stream to the nerves is restricted leaving the nerves without blood stream and they get harmed or kick the bucket subsequently. The symptoms of nerve damage are numbness burning, aching of feet and lower extremities [3]. There are two major types of diabetes type I and type II. Type I is basically diagnosed in children which is usually known as Juvenile diabetes and type II is most common form of diabetes. A patient with type II diabetes do not require insulin cure to remain alive, although up to 20% are treated with insulin to control blood glucose level [4]. Diabetes mellitus is a degenerative infection described by either absence of insulin or a protection from insulin, a hormone which is pivotal

for digestion of glucose. In a sound individual, the pancreas produces insulin to help process sugar in the blood and keep up blood glucose (sugar) levels inside their ordinary range. Diabetics can't deliver insulin or are impervious to insulin, and subsequently can't expel glucose from the circulatory system. Regardless of whether there is insufficient insulin or insulin protection, glucose levels in the blood increment and cause extreme medical issues. There are two noteworthy sorts of diabetes. Sort 1, or insulin-subordinate adolescent diabetics, is hereditary in cause and is described by the body's powerlessness to create insulin, and the subsequent development of glucose in the blood. It normally happens amid pubescence or immaturity however can happen amid adulthood. Side effects incorporate outrageous craving and thirst, visit or over the top pee, and weight reduction. Definitive impacts of diabetes incorporate coronary illness, kidney malady, malignancy, hypertension, gangrene, diseases, visual deficiency, strokes, and demise.

Type II, or non-insulin-subordinate grown-up beginning diabetes, is the more typical frame and is portrayed by the body's protection from insulin. Around 90% of diabetes are type II and 80% of them are overweight when analyzed, for the most part amid middle age. Most stout diabetics have lifted insulin levels, however it doesn't control their glucose on account of the deficiency in insulin receptor cells and insulin protection. Stoutness and overabundance calories make a protection from insulin - that is, the pancreas keeps on creating insulin in light of blood glucose, yet the body's cells oppose the activity of insulin. The mix of heftiness and high glucose prompt a lessening in the quantity of insulin receptors, destinations to which insulin connects to start transformation of glucose to glycogen or fat for capacity. Weight reduction and diminished caloric admission cause an expansion in the quantity of receptor cells prompting more proficient insulin digestion [5].

Indications of Type II diabetes are the same as side effects of Type I, however weight reduction is once in a while experienced in a Type II diabetic without an adjustment in count calories. Dissimilar to Type I diabetes, which for the most part requires normal insulin infusions, Type II diabetes can ordinarily be controlled by regular strategies, including diet, weight control, hormonal adjusting, compound treatment and home grown supplements. By and large, doctors consider a fasting plasma glucose level over 140 mg/gI as exorbitant.

II. DIABETES DATA AND PRE-PROCESSING

Diabetes prediction using machine learning requires high-quality data and effective pre-processing techniques to ensure optimal model performance. Common datasets used for this task include the PIMA Indian Diabetes Dataset (PIDD), electronic health records (EHRs), and data from wearable devices, which contain a mix of clinical, demographic, and lifestyle variables.

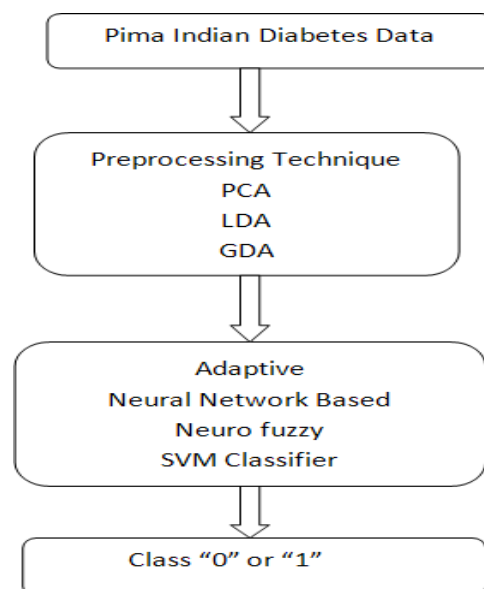


Fig. 1: Data pre-processing of Pima Indian data set

Pre-processing steps are essential for cleaning and preparing the data. These steps involve handling missing values through imputation techniques, detecting and managing outliers, and ensuring consistency in data formatting. Feature scaling and normalization standardize the data, while encoding methods like one-hot encoding transform categorical variables into numerical formats. Addressing class imbalance is crucial, often achieved through oversampling, undersampling, or weighted loss functions in algorithms like XGBoost. Additionally, feature selection techniques such as Recursive Feature Elimination (RFE) or SHapley Additive exPlanations (SHAP) are employed to identify the most influential predictors. Finally, splitting the data into training, validation, and testing subsets ensures the model generalizes effectively to unseen data. This comprehensive pre-processing pipeline is pivotal for developing accurate and reliable diabetes prediction models.

III. PROPOSED FRAMEWORK

The XGBoost (eXtreme Gradient Boosting) classifier is a powerful and efficient machine learning algorithm widely used for classification tasks, including diabetes risk prediction. Built on the principles of gradient boosting, XGBoost enhances weak learners by sequentially adding decision trees to minimize prediction errors. It is known for its high performance, scalability, and ability to handle large datasets with missing values and noisy data. XGBoost incorporates features like regularization to prevent overfitting, parallel tree construction for computational efficiency, and built-in cross-validation for model optimization. The algorithm supports various evaluation metrics, making it adaptable to imbalanced datasets by applying class weights or customized loss functions. Additionally, XGBoost offers interpretability through feature importance scores, enabling the identification of key predictors influencing model decisions. Its flexibility, accuracy, and robustness make the XGBoost classifier a preferred choice for predictive modeling in healthcare, providing reliable insights for early diabetes detection and management.

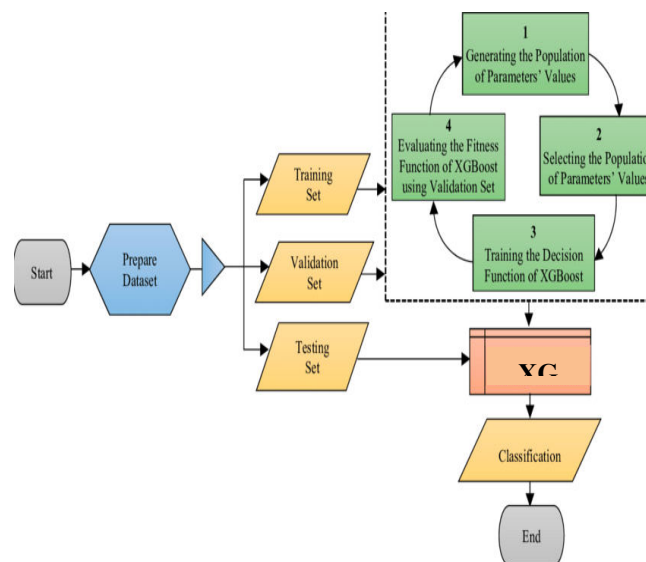


Fig. 2: Flow chart of Proposed Algorithm

Dataset was divided into two datasets (70%/30%, training/testing) to avoid any bias in training and testing. Of the data, 70% was used to train the ML model, and the remaining 30% was used for testing the performance of the proposed activity classification system. The expressions to calculate precision and recall are provided in Equations (2) and (3).

Precision provides a measure of how accurate your model is in predicting the actual positives out of the total positives predicted by your system. Recall provides the number of actual positives captured by our model by classifying these as true positive. F-measure can provide a balance between precision and recall, and it is preferred over accuracy where data is unbalanced.

Algorithm steps:

Input: $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}, L(y, O(x))$

Where: $(y, O(x))$ is the approximate loss function.

Begin

Initialize: $(x) = \underset{w}{\operatorname{argmin}} \sum_{i=1}^n L(y_i, w)$

for $m=1:M$

$r_{im} = - \frac{\partial L(y_i, O(x_i))}{\partial O(x_i)}$

Train weak learner $C_m(x)$ on training data

Calculate w : $w_m = \operatorname{argmin} \sum_{i=1}^N L(y_i, O_{m-1}(x_i) + wC_m(x_i))$

Update : $O_m(x) = O_{m-1}(x) + wC_m(x)$

End for

End

Output: $O_m(x)$

IV. EXPERIMENTAL RESULTS

Data Set

In this section describes the detailed analysis of experimental works carried out for our proposed model. The computational complexity of the proposed algorithm may change for different datasets depending on its size. The parametric values may vary accordingly.

Dataset description:

The source of Pima Indians dataset diabetes dataset on which the experiment is performed is UCI machine learning repository [12] with 768 data instances and 9 attributes. All patients in this dataset are Pima Indians women whose age is at least 21 years old and living near Phoenix , Arizona which denotes either “0” or “1”, where ‘0’ is tested as negative and ‘1’ is tested as positive for diabetes.

Table 1: Description of benchmark dataset for diabetic for pima Indians

Datasets	No. of features	No. of classes	No. of patterns
Pima India Diabetic Dataset	9	2	686

Table 2: Description of parameter used for FFFNN

Datasets	Description	Considered value/Size
N	Number of input vector	768
D	Desired output vector	768
M	Number of hidden neurons	15
W	Weight vector	150
N	Number of input neuron	9
X	Input vector	768x9

Software

For implementation, we use MATLAB software with version 7.10.0. The coding for Classification using modified PSO-FFNN is executed in the command window. The operating system used is windows operating systems with 2 GB RAM. The results tabulated in the table Table 6 are carried out in MATLAB.

Table 3: Description of Diabetic Data Set

Data Set	No. of Attributes	Feature Set
Diabetic	9	No. of times pregnant Plasma glucose concentration Diastolic blood pressure Triceps skin fold thickness Serum insulin Body mass index Diabetes pedigree function Age of patient Class '0' or '1'

Parameter details:- The different significant parameter used for FFNN are center, spread and weight. The different symbols used for FFNN, PSO and IPSO are described in table 2 and table 3.

Evaluation metrics: Generally, the evaluation of a classification problem is based on a matrix called as a confusion matrix with the number of testing samples correctly classified and incorrectly classified represented as follows. So, the accuracy can be measured according to Eq. 1

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{1}$$

For a binary classification problem, the other measures include Precision, Sensitivity or Recall and Specificity. The formula to derive these measures is given in Eq. 6 and Eq. 7.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

In these relations ((1),(2) and (3) formula) TP means the number of samples that are healthy and properly diagnosed. FP indicates the number of samples that are healthy and have been diagnosed wrongly. FN indicates the number of samples that were sick but healthy wrongly diagnosed. TN contains a number of examples that have been patient and the patient is properly diagnosed.

Here is the CM for Proposed Algorithm is representing in fig. 3. In proposed calculation we utilized an outfit of Slope supporting to accomplish a precision of 92.18%. The Larger part vote-based model as shown which contains L.R., NV, R.F.C., K.N.N., D.T. what's more, SVM classifiers gave an exactness of 82.11%, 81.65%, 82.11%, 74.77%, 81.19% and 79.35% for diabetes sickness dataset.

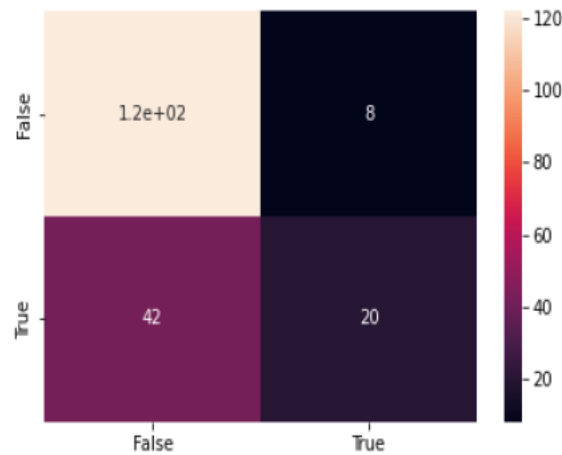


Figure 3: Confusion Matrix for Proposed Algorithm

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 4, 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 4.

Table 4: Assessment of Different Classification Methods

Sr. No.	Algorithm	Accuracy
1	LR	0.8211
2	GNV	0.8165
3	RFC	0.8211
4	K-NN	0.7477
5	DT	0.8119
6	SVM	0.7935
7	Proposed Algorithm	0.9218

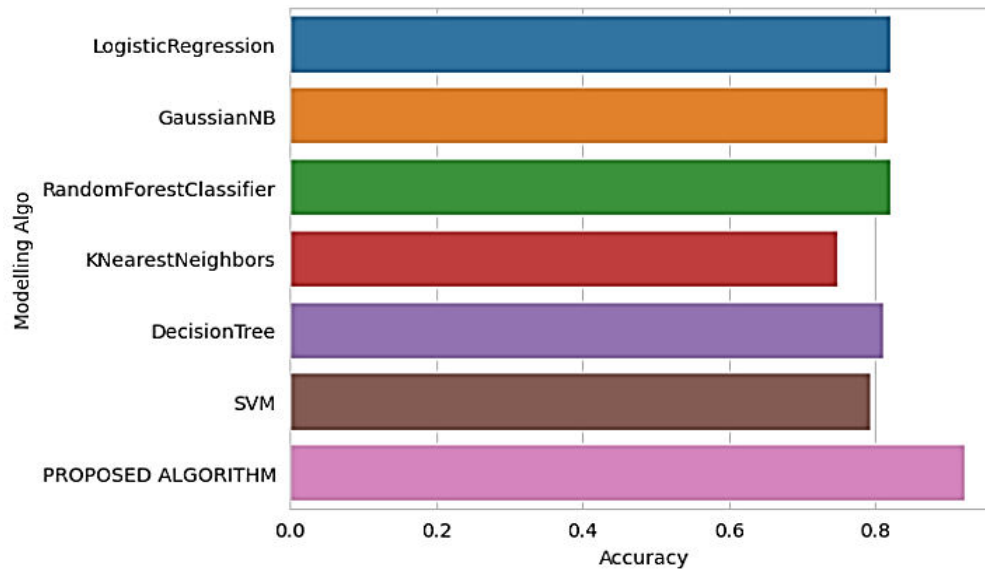


Figure 4: Bar Graph of the Different Classification Methods

V. CONCLUSION

The XGBoost classifier has proven to be a highly effective and reliable tool for predicting diabetes risk, offering significant advantages in handling complex datasets and imbalanced classes. Its robust feature set, including regularization, parallel processing, and interpretability, makes it particularly well-suited for healthcare applications where accuracy and transparency are paramount. By leveraging clinical, demographic, and lifestyle data, the XGBoost classifier can provide accurate predictions, enabling early intervention and personalized healthcare strategies.

This study highlights the potential of integrating XGBoost-based models into healthcare systems to improve decision-making, optimize resource allocation, and reduce costs associated with diabetes management. While challenges such as data quality and interpretability remain, advancements in feature engineering, algorithm optimization, and multi-modal data integration can further enhance the utility of XGBoost in predictive healthcare. As machine learning continues to evolve, the application of techniques like XGBoost offers a transformative approach to proactive disease management, fostering a healthier and more sustainable future.

REFERENCES

1. 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 Volume 9, Issue 01, January 2020.
2. 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 2018.
3. Rao G.A., Syamala K., Kishore P.V.V., Sastry A.S.C.S. ., "Deep convolutional neural networks for sign language recognition", 2018, International Journal of Engineering and Technology(UAE) ,Vol: 7, Issue 5, pp: 62 to 70.
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, no. c, pp. 8869–8879, 2017.
5. 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.
6. 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.
7. Reddy S.S., Suman M., Prakash K.N. ., "Micro aneurysms detection using artificial neural networks", 2018, Lecture Notes in Electrical Engineering ,Vol: 434 ,Issue 3, pp: 409 to 417.

8. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., “Machine Learning and Data Mining Methods in Diabetes Research”, Computational and Structural Biotechnology Journal 15, 104–116, 2017.
9. Majid Ghonji Feshki and Omid Sojoodi Shijan, “Improving the Heart Disease Diagnosis by Evolutionary Algorithm of PSO and Feed Forward Neural Network”, International paper on IEEE 2016.
10. L. Hermawanti, "Combining of Backward Elimination and Naive Bayes Algorithm To Diagnose Breast Cancer", Momentum, vol. 11, no. 1, pp. 42-45, 2015.
11. O.S. Soliman, E. Elhamd, “Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine”, IEEE 2014.
12. K. Saxena, Z. Khan, S. Singh, "Diagnosis of Diabetes Mellitus using K Nearest Neighbour Algorithm", International Journal of Computer Science Trends and Technology (IJCST), 2014.
13. L. Hermawanti, S.G. Rabiha, "Combining of Backward Elimination and K-Nearest Neighbor Algorithms To Diagnose Heart Disease", Prosiding SNST Ke-5 Fakultas Teknik Universitas Wahid Hasyim, pp. 1-5, 2014.
14. R.A. Vinarti, W. Anggraeni, "Identification of Prediction Factor Diagnosis of Breast Cancer Rates with Stepwise Binary Logistic Regression Method", Jurnal Informatik, vol. 12, no. 2, pp. 70-76, November 2014.
15. Muhammad Waqar Aslam, Zhechen Zhu and Asoke Kumar Nandi, “Feature generation programming with comparative partner selection for diabetes classification”, “Expert Systems with Applications”, 5402-5412, IEEE 2013.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details