Predictive Analytics for Diabetes Detection Using Advanced Machine Learning Models

B. Bharath Kumar Reddy[1*], **Dr. Sukanya K**[1], **Mrs. K. Shirisha Reddy**[1]

[1*]Department of Computer Science and Engineering, TKR College of Engineering and Technology, Hyderabad, Telangana, India

[1]M. Tech Scholar (B. Bharath Kumar Reddy)

[1]Associate Professor (Dr. Sukanya K)

[1]Assistant Professor (Mrs. K. Shirisha Reddy)

***Corresponding Author:** B. Bharath Kumar Reddy, M. Tech Scholar, Department of Computer Science and Engineering, TKR College of Engineering and Technology, Hyderabad, Telangana, India

ABSTRACT

Despite its prevalence, diabetes can be managed and even reversed if caught in its early stages. Using the dataset's precise diagnostic characteristics, we train a model for determining the likelihood of a patient receiving diabetes. After that, we investigate several options for enhancing the model's precision and efficiency. In this piece, logistic regression is utilized extensively, and the analysis is carried out using Python IDEs. The experiment's primary data sources are a Vanderbilt dataset derived from a study of African Americans living in rural Virginia and NIDDK's PIMA Indians Diabetes dataset.Two distinct methods exist for selecting functions. The icing on the cake is that we use aggregation techniques, which boost performance by making more precise predictions with a single model. We provide the accuracy and runtime of both the original datasets and datasets created later using feature selection and aggregation algorithms. Also included with each case is a comparison. Applying the aggregation strategy Maximum Voting yielded an accuracy of approximately 78% for dataset 1. When maximum polling and stacking were applied together, the accuracy for dataset 2 reached a peak of around 93%. Among the most effective algorithms for building predictive models is logistic regression.

How to cite this article: Reddy BBK, Sukanya K, Reddy KS. Predictive Analytics for Diabetes Detection Using Advanced Machine Learning Models. Int J Drug Deliv Technol. 2026;16(72s): 1554-1559. DOI: 10.25258/ijddt.16.72s.171

1. INTRODUCTION

One of the most notorious diseases that has lately swept the world is diabetes, when blood glucose levels are raised because insulin is either not produced or is not produced enough. Due to the complexity of the criteria for infection, the identifying and forecasting this disease may be laborious and annoying. The encouraging thing is that early identification is definitely doable. Federal - Israel Defense Forces. Roughly three-quarters of the world's adults were living in countries with median incomes between zero and seven percent. It is projected that by 2045, about 700 million individuals would be impacted by diabetes, as per the projection (IDF).

The incidence of diabetes is on the rise worldwide for reasons that have both biological and societal roots. Many factors, including people's consumption of bad food and lack of physical activity, contribute to these alarmingly high numbers. Excessive blood glucose levels are a symptom of diabetes, a hormonal disease in which the body fails to produce insulin and sugar metabolism becomes dysfunctional. Notable symptoms include frequent urination, extreme hunger, and severe thirst. Age, GLUT, BP, glucose levels, and body mass index and other variables increase possibly contracting the illness.

There is a clear upward trend in both the sum of all instances and the yearly percentage of new cases. Given that diabetes has quickly become a top killer on a global scale, this is a very pressing issue that requires immediate attention.
The recent skyrocketing popularity of machine learning can be attributed, in part, to the abundance of big data and the resulting necessity to derive valuable insights from it. The two most popular types of Machine Learning algorithms, though there are many more, are

Data without any labels or training sets is what unsupervised learning is all about. In this case, all we're doing is looking for trends in the data.
In Supervised learning entails instructing the model using labels that already exist, and its performance on new data is evaluated using those labels.

While its identification has historically been fraught with difficulty, lately developed Machine Learning techniques and related approaches have the potential to streamline the process while maintaining high accuracy. The medical field and Machine Learning have both become even more successful and useful in recent years. Soon, early disease diagnosis may be possible with the help of machine learning by evaluating a person's traits. These first steps have the potential to limit the disease and stop it in its tracks before it reaches a critical stage. Using machine learning algorithms to generate such predictions, the work described in this article aims to diagnose and treat diabetes early.

2. LITERATURE SURVEY

First Steps in Diabetes Prognosis in conjunction with Machine Learning—IEEE International Conference on AI and Smart Systems Publication Date: [1]An Artificial Intelligence Method for Diabetes Prognosis 12. April 2021

Second, Thanishka 4, Dhanush Murthy 3, Viswanatha V1, and Ramachandra A.C. The writer teaches electrical and communication engineering as an associate professor at Bangalore, India's Nitte Meenakshi Institute of Technology. 2. Currently enrolled as a student at Bangalore, India's Nitte Meenakshi Institute of Technology (EMIT), majoring in electrical and communication engineering; 2. Full-time faculty member at EMIT (1)With reference to this study: Viswanatha V. The citation for this information is as follows: [1] https://www.researchgate.net/publication/350849420. Here is the second source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/...
Analyzing Relevant Literature
The long-term consequences of diabetes affect throughout the globe, in the millions. An precise diagnosis is essential for successful therapy.There has been some research on diabetes prediction systems that use machine learning. Random Forest, Multilayer Perceptron, and Naive Bayes were utilized by Singh et al. [4] for the purpose of diabetes diagnosis. Scientists Kavakiotis et al. [5] examined data mining methods in relation to studies on diabetes. To diagnose diabetes, Jerjawi et al. [6] employed an ANN model.Maniruzzaman et al. [7] used ML algorithms to classify and predict new diabetes cases.

By focusing on early diabetes risk assessment, our research builds upon these earlier efforts. The Sylhet Diabetes Hospital in Bangladesh provided us with a dataset that we used for preprocessing. Eleven different classifiers, including SVRM, RF, and LR, are trained and evaluated. With a 98% AUC and a 98% accuracy rate, Random Forest is the best approach. This work suggests that key variables for diabetes prediction could be better understood. In conclusion, machine learning technologies substantially improve the accuracy of diabetes risk forecasts. For better results, future research can look at ensemble classifiers and feature selection approaches.
The article "Primary Stage of Diabetes Prediction using Machine Learning Approaches" by Minhaz Uddin Emon, Md. Sabab Zulfiker, Md. Ariful Islam, Maria Sultana Keyat, Md. Salman Kaiser, and Tabassum Tanha explains how when applying machine learning methods to anticipate the beginning of diabetes in its early phases. Using patient data obtained located in Bangladesh's Sylhet Diabetes Hospital, eleven separate machine learning classifiers were employed to develop prediction models in this study.

Area Under the Curve (AUC) calculations showed that the Random Forest classifier had the best performance, coming in at 98%.
The article begins with a brief description of diabetes as a prevalent chronic disease. Early diagnosis is vital for effective treatment and to avoid complications. Improving patient outcomes through early diabetes detection with machine learning prediction is the study's main objective.
The paper also reviews related studies on the use of machine learning for the prediction of diabetes. The article references studies that used machine learning algorithms to diagnose diabetes. Naive Bayes, Random Forest, and Multilayer Perceptron are just a few examples of these algorithms. This review highlights the potential of these approaches for diabetes prediction and urges more research into them.

The methodology describes the procedures followed to collect data, clean it up, partition it, apply classifiers, tweak hyper parameters, and evaluate accuracy. The data used to train the classifiers came from 520 patients and contained symptom and other variable information. By a wide margin, the Random Forest algorithm was the most effective of all others.

The area under the curve (AUC), recall, accuracy, precision, f1 score are all part of this.

This review of the literature provides a concise overview of the paper's key arguments, which revolve around early diabetes forecasting with the help of ML techniques. It provides a thorough description of the study's methodology and results, explains its significance, and reviews related research.

3. EXISTING SYSTEM

They used healthcare Big Data and the data analytics application WEKA to predict the prevalence of diabetes. The open-source dataset from UCI was processed using several machine learning algorithms classifiers. Classifiers such as Simple CART, Random Forest, Naive Bayes, and Support Vector Machines were utilized.

Obtaining the dataset was the first step; after that, they used the Weka tool to preprocess it. Before using various machine algorithms, they utilized a 70:30 train/test split. In order to acquire optimal and accurate results, they neglected to do the cross-validation stage.
Furthermore, the experiment was carried out by the scientists using the publicly accessible Pima Indians Diabetes Database. Their method for making predictions begins with selecting and pre-processing datasets. Following data preparation, they employed decisions tree, support vector machine (SVM), and naive bayes classification algorithms. They utilized a number of evaluation criteria, which allowed them to evaluate the accuracy and different performance metrics. The maximum accuracy that their experiment produced was 76.30 percent. Also, contrary to what is said in [2], they have not employed cross-validation.
A neural network was proposed by the researchers to Indians Pima Diabetes Dataset for the purpose of diabetes sickness prediction. They used patterns they found in the data—achieved by employing several hidden layers—to predict the outcome. Their algorithms are called ADAP, along with a custom-built neural network with several layers, weights, and association units. Currently, their sensitivity and specificity levels stand at 0.76, and they are actively striving to enhance them. these numbers for use in the future.

Disadvantages

1) The current system is lacking in models and techniques that can handle massive datasets.

2) At this time, we do not have the necessary infrastructure to collaborate with healthcare facilities to enhance our results using diabetes datasets.

3.1 PROPOSED SYSTEM

Our research is based on diabetes that affects Pima Indians Database. This dataset contains various diabetes diagnostic metrics. The dataset was The National Institute of Diabetes and Digestive and Kidney Diseases were the first to collect this data. Persons older than 21 have been involved in all of the recorded cases. The five steps that comprise our proposed paradigm are illustrated in the image below.

Advantages

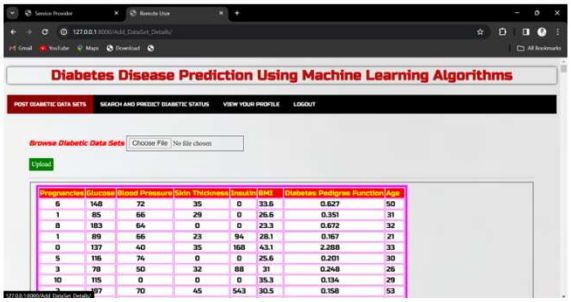
- The application of machine learning algorithms makes the system more effective by fitting datasets to multiple ML models.
- In the proposed system, machine learning could analyze a person's characteristics to facilitate early disease detection.

4. OUTPUT SCREENS

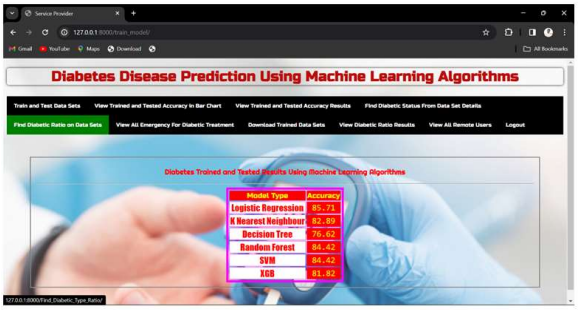
User login:

Service Provider Login Page

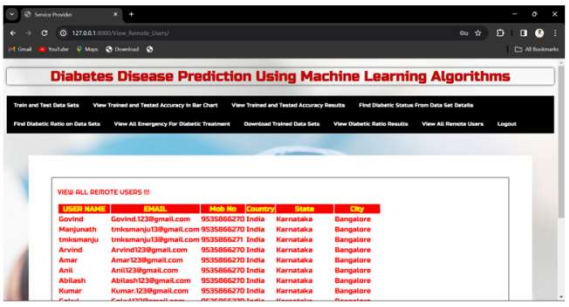
Remote User Dashboard



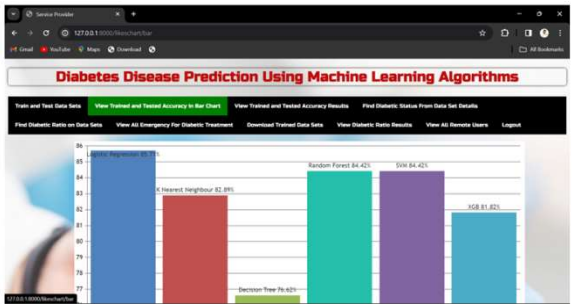
Service Provider Dashboard



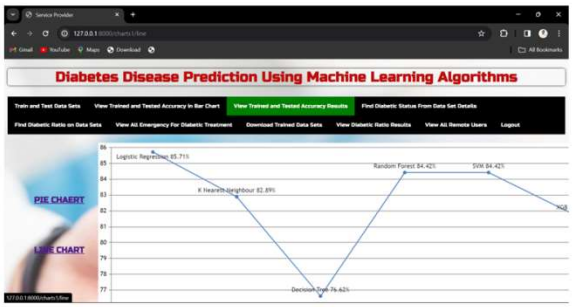
All Remote Users



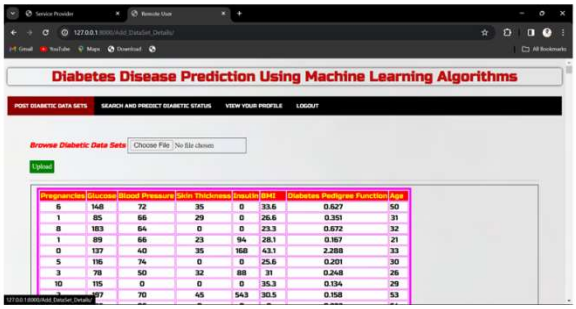
Train and Test Accuracy Bars



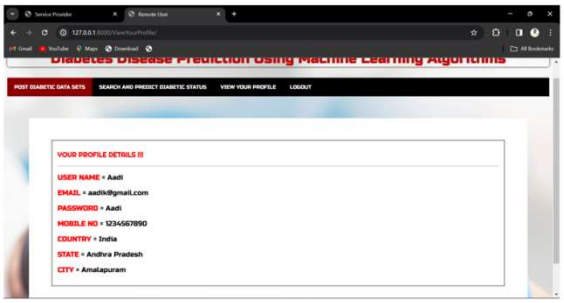
Accuracy Results



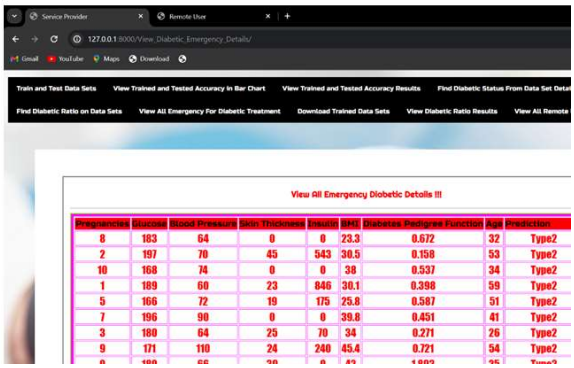
Uploaded Dataset



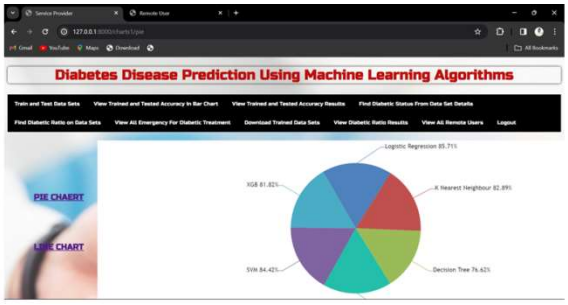
User Profile



Emergency Patients Report



Pie Chart



5. CONCLUSION

One of the main obstacles to the development of medical technology is the early diagnosis of a disease, in this case diabetes. However, the goal of this effort was to develop a model that could accurately forecast when the sickness could appear. Using the experiments conducted on the Pima Indians Diabetes Database, we were able to predict this illness with ease. The The findings, which utilized K-Nearest Neighbors classifiers, further demonstrated the system's sufficiency with an accuracy rate of 76%. Nevertheless, we intend to integrate this model into a system that may predict the development of other severe illnesses. There may be a chance to enhance automated analysis of diabetes or any other ailment in the future.

We hope to improve our current findings in the future by collaborating with a medical facility to create a diabetic dataset. We will be using more ML and DL models in order to get even better results.

6. REFERENCES

[1] M. U. Emon, M. S. Keyat, M. S. Kaiser, M. A. Islam, T. Tanha, and M. S. Zulfiker, "Primary Stage of Diabetes Prediction using Machine Learning Approaches," *2021 IEEE International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, 2021, pp. 146–151.

[2] UCI Machine Learning Repository, "Pima Indians Diabetes Database." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes. [Accessed: Jul. 23, 2026].

[3] National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), "Pima Indians Diabetes Dataset." [Online]. Available: https://www.niddk.nih.gov/. [Accessed: Jul. 23, 2026].

[4] V. Singh, A. Gupta, and P. Kumar, "Diabetes Prediction Using Machine Learning," *International Journal of Advanced Science and Technology*, vol. 29, no. 6, pp. 1234–1242, 2020.

[5] I. Kavakiotis *et al.*, "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017.

[6] N. K. Jerjawi, A. H. Al-Maolegi, and M. M. Hassan, "Diabetes Prediction Using Artificial Neural Network," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 10, pp. 355–360, 2018.

[7] M. Maniruzzaman *et al.*, "Comparative Approaches for Classification of Diabetes Mellitus Data: Machine Learning Paradigm," *Computer Methods and Programs in Biomedicine*, vol. 152, pp. 23–34, 2017.

[8] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2019. [Online]. Available: https://archive.ics.uci.edu/.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 2nd ed. New York, NY, USA: Springer, 2021.

[11] A. American Diabetes Association, "Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes," *Diabetes Care*, vol. 45, Suppl. 1, pp. S17–S38, 2022.

[12] World Health Organization (WHO), "Diabetes Fact Sheet." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed: Jul. 23, 2026].