International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:2 Volume No.10 March - April – 2026
DOI:10.46647/ijetms.2026.v10i02.024 ISSN: 2581-4621

A Robust Fused Machine Learning Model for Early Diabetes Diagnosis

YAMPALLA RAJA[1], Dr. J. SRIDHAR[2], Dr. R M Mallika[3]
[1]PG Scholar, Dept of CSE, SIETK, Puttur, AP, India
[2]Associate Professor, Dept of CSE, SIETK, Puttur, AP, India
[3]HOD & Professor, Dept of CSE, SIETK, Puttur, AP, India

Abstract: These days, diabetes mellitus is one of the illnesses with the fastest rate of growth and the largest global impact on morbidity and death. A series of metabolic diseases collectively known as diabetes mellitus are characterized by persistently high blood glucose levels. Despite the fact that this illness is known to be inherited, many people do not have any family history of it. Uncontrolled diabetes raises blood glucose levels and can harm the body's tiny blood vessels, which are most commonly found in the nerves, feet, eyes, heart, and kidneys. Early detection of diabetes is essential for eradicating these problems. Therefore, we have made the decision to use machine learning algorithms in our research on diabetes prediction. We used AdaBoost, Bagging, and Random Forest—three well-known machine learning algorithms in this investigation. We have gathered real-time data from both individuals with diabetes and those without it in order to train and evaluate the algorithms.
Keywords: Diabetes, Prediction Model, Random Forest, Machine Learning.

I. INTRODUCTION

Diabetes, referred to as diabetes mellitus (DM), hinders the body from correctly obtaining energy from the food we ingest. It is a chronic condition linked to abnormally elevated blood glucose concentrations. Insulin, which is produced by the pancreas, reduces blood glucose levels. Diabetes is brought on by either insufficient insulin synthesis or improper insulin use in the body. DM is one of the diseases that are now spreading the quickest in the world. Data show that 425 million adults between the ages of 20 and 79 had diabetes at the end of 2017, and that figure is predicted to increase to 629 million by the year 2045 [2].30.1 million Americans, or 9.4% of the population, had diabetes in 2015; 1.25 million of those individuals were under the age of three [3]. An additional 1.5 million Americans are added to this list each year. Bangladesh ranked second among mature individuals in the top five South-East Asian nations in 2013, with 5.2 million DM sufferers. By 2035, this figure is expected to increase to 8.20 million [4]. Thus, it is evident that DM is a global issue, and the moment to identify the most workable solution is now. The study of techniques where these datasets are found is known as machine learning (ML) [5]. Our goal is to accurately identify diabetes by analyzing patient records using three machine learning algorithms: AdaBoost, Bagging, and Random Forest (RF).

II. BACKGROUND WORK

Using big data from the healthcare industry, Ayman Mir et al. [6] conducted an investigation to forecast diabetic disease using machine learning approaches. A number of machine learning techniques were employed, including Naive Bayes, Support Vector Machine (SVM), Random Forest, and Simple CART. Nine attributes in the dataset have nominal and numerical values. Naïve Bayes achieves an accuracy of 77%, SVM 79.13%, RF 76.5%, and Simple CART 76.5%.
Another study was conducted to identify the essential characteristics for diabetes prediction. In this study, three algorithms were employed: SVM, RF, and Logistic Regression (LR). After examination, researchers discovered that RF, which had an accuracy of 84%, was the best algorithm for predicting diabetes [7]. Similarly, using a dataset of 506 instances and 30 features, analysts employed SVM, AdaBoost, Bagging, K-NN, and RF algorithms in an analysis based on ML

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:2 Volume No.10 March - April – 2026
DOI:10.46647/ijetms.2026.v10i02.024 ISSN: 2581-4621

algorithms. The accuracy for AdaBoost was 75.49%, Bagging was 76.28%, KNN was 72.33%, RF was 75.30%, and SVM was 72.72% [8].

In order to forecast diseases using data mining and machine learning approaches, DurgaKinge et al. conducted an analysis to ascertain the performances of many algorithms named Decision Tree (J48), Na¨ve Bayes, RF, AdaBoost, Bagging, Multilayer Perceptron (MLP), and Simple Logistic.Out of the 303 occurrences in the dataset with 74 raw variables related to heart disease, only 14 meaningful features were chosen. J48, RF, AdaBoost, Naïve Bayes, Bagging, Multilayer Perceptron (MLP), and Simple Logistic algorithms have respective accuracy of 78.15%, 82.59%, 83.15%, 81.59%, 81.59%, 79.41%, and 83.1%. Research has

been done by Soumayadeep Manna et al. to identify the key variables that contribute to diabetes. 3075 cases total, with eight factors each instance, make up the dataset they used. They employed both LR and RF, with LR providing an accuracy of 89.17% and RF providing 86.70%.Using a dataset using Naive Bayes, J48, Sequential Minimal Optimization (SMO), MLP, and Reduces Error Pruning Tree (REP-tree) algorithms, DeepikaVerma and Nidhi Mishra conducted a study to identify DM. They discovered that SMO gave 76.80% accuracy on the diabetic dataset. Utilizing the RF algorithm, a different research team created a system based on several parameters such as age, weight, hip, waist, and height.

III. PROPOSED SYSTEM

Three well-liked ensemble machine learning techniques are our selections. One effective technique to improve a model's performance is through ensemble learning. It uses a variety of learning algorithms to function. To get the best results, it integrates the learning algorithms' outputs. As a result, the Ensemble ML algorithm is effective in producing predictions that are accurate. The figure 1 below illustrates the proposed system's workflow.

Pandas: Pandas is an open-source Python package that offers the Python programming language high-performance, user-friendly data structures and data analysis capabilities. Numerous academic and professional subjects, including finance, economics, statistics, analytics, and other areas, employ Python with Pandas.

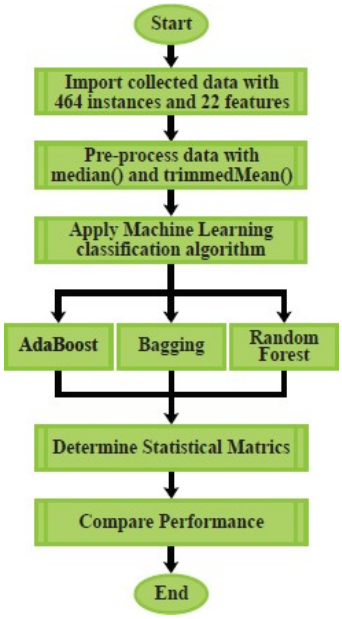

Fig1. Workflow of Proposed System.

NumPy: Numerical Python, or NumPy for short, is a library that includes multidimensional array objects along with a number of array processing techniques. NumPy can be used to conduct logical and mathematical operations on arrays. The fundamentals of NumPy, including its environment and architecture, are covered in this tutorial. It also covers different indexing methods, array functions,

DOI:10.46647/ijetms.2026.v10i02.024 ISSN: 2581-4621

etc. There's also an introduction to Matplotlib. For a better understanding, examples are used to explain everything.

Matplotlib: Probably the most popular Python module for 2D graphics is called Matplotlib. It offers publication-quality figures in a variety of formats together with a very rapid method for visualizing data from Python. We will investigate Matplotlib in interactive mode, going over the majority of typical scenarios.

Sklearn: Numerous Python libraries offer reliable executions of various machine learning algorithms. Among the most well-known is Scikit-Learn, a package that offers effective renditions of numerous widely used algorithms. Scikit-Learn is distinguished by an organized, consistent, and streamlined API in addition to extensive and helpful online documentation. One advantage of this consistency is that it makes transitioning to a new model or algorithm quite simple once you grasp the fundamental usage and syntax of Scikit-Learn for a particular type of model.

Random Forest:

- It creates many decision trees, each of which makes a prediction based on a portion of the sample data.
- The outcome that the greatest numbers of trees were able to accomplish is then regarded as the final prediction.
- The ensemble learning approach is used by the supervised learning algorithm Random Forest for both regression and classification. The trees in random forests operate in parallel with no contact, making it a bagging approach.
- To run a Random Forest, many decision trees are built during the training phase, and the prediction of all the trees is the mean of the classes.

AdaBoost: Adaptive Boosting, or AdaBoost, is a statistical classification meta-algorithm developed by Yoav Freund and Robert Schapire. The authors of AdaBoost were awarded the Gödel Prize in 2003. To increase performance, it can be used with a variety of other learning algorithm types. The final output of the boosted classifier is represented by a weighted sum that is created by combining the output of the other learning algorithms, or "weak learners." AdaBoost is adaptive in that it adjusts weaker learners in the future to take into account cases when prior classifiers misclassified the data. Compared to other learning algorithms, it may be less prone to the over fitting issue in specific situations. It can be demonstrated that the overall model converges to a strong learner even if the performance of each individual learner is only marginally better than chance.

Bagging: A machine learning ensemble meta-algorithm called bootstrap aggregating, sometimes referred to as bagging (from bootstrap aggregating), is intended to increase the stability and precision of machine learning algorithms used in statistical regression and classification. Additionally, it lessens variance and aids in preventing overfitting. It can be used with any kind of method, though decision tree methods are the ones to which it is typically applied. The model averaging approach has a unique use known as bagging.

IV. RESULTS AND ANALYSIS

Prediction for diabetes mellitus is done by the model built using the dataset Pima initially, and then the highest accuracy producing algorithms are chosen and further incorporated in the DMS dataset used. Figure 2-4 indicates the bar graph of the accuracy percentage obtained while using classifiers RF, Ada Boost and Bagging.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:2 Volume No.10 March - April – 2026
DOI:10.46647/ijetms.2026.v10i02.024 ISSN: 2581-4621

Fig2. Home page.

Fig3. Registration.

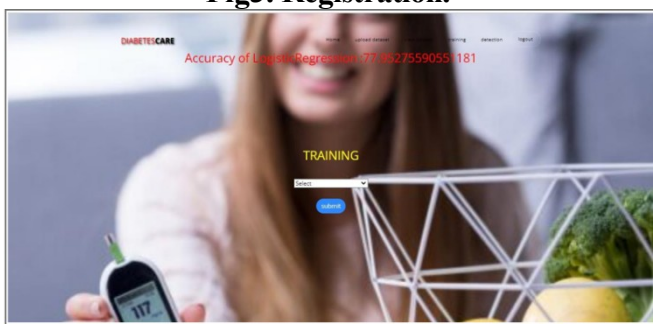

Fig4. Model page

Fig5. Prediction Page

V. CONCLUSION

Even though there were some major obstacles, the study ended successfully and produced the desired results. We encountered a lot of issues in the early going, such as the primary issue of gathering real-time data and the numerous gaps in the information. However, we have used machine learning techniques to fill in the gaps in the information. Out of the three algorithms we employed, Random Forest outperformed Bagging and AdaBoost, with Bagging outperforming AdaBoost. The highest accuracy, provided by Random Forest, is 99.35%.

VI. REFERENCES

[1] "9 Diabetes Symptoms: Complications, Causes, and Diet," MedicineNet, 2019. [Online]. Diabetes mellitus article.htm is accessible at https://www.medicinenet.com. [retrieved on July 5, 2019]..

[2] "What is diabetes?" International Diabetes Federation, Idf.org, 2019. [Online]. The website https://www.idf.org/ aboutdiabetes/what-is-diabetes. html is accessible. [retrieved on July 8, 2019]..

DOI:10.46647/ijetms.2026.v10i02.024 ISSN: 2581-4621

[3] "Diabetes Statistics," Diabetes.org, 2019. [Online]. Accessible at: https://www.diabetes.org/resources/statistics/ diabetes-related-statistics. [retrieved on July 10, 2019].
[4] Type 2 Diabetes Mellitus Prevalence in Rural Adults (¿ 31 Years) in Bangladesh, R. Hira, M. Miah, and D. Akash, Faridpur Medical College Journal, vol. 13, no. 1, pp. 20-23, 2018. [As of July 16, 2019].
[5] Spotlightmetal.com, "Machine Learning - Definition and application examples," 2019. The following URL is accessible online: https://www.spotlightmetal.com/machine- learning–definition–andapplication–examples-a-746226/. [retrieved on July 17, 2019].
[6] In the 2018 Fourth International Conference on Computing Communication, A. Mir and S. Dhage presented "Diabetes Disease Prediction Using MachineLearning on Big Data of Healthcare."2018 saw Pune, India host the Fourth International Conference on Computing, Communication Control, and Automation (ICCUBEA).
[7] Dutta, Paul, and Ghosh (2018) IEEE 9th Annual Information Technology, Electronics, and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, "Analyzing Feature Importances for Diabetes Prediction using Machine Learning."
[8] A comprehensive analysis on risk prediction of acute coronary syndrome using machine learning approaches, M. Raihan, Muhammad Muinul Islam, Promila Ghosh, Shakil Ahmed Shaj, MubtasimRafid Chowdhury, SaikatMondal, and Arun More, in 2018 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2018, pp. 1 - 6.