OPEN ACCESS

Volume: 4

Issue: 1

Month: January

Year: 2025

ISSN: 2583-7117

Published: 06-01.2025

Citation:

Shivani Sahu, Jayshree Boaddh, Meena Deshbhratar, Combining SVM and Gradient Boosting for Enhanced Accuracy in Diabetes Diagnosis, International Journal of Innovations In Science Engineering And Management, vol. 4, no. 1, 2025, pp.08-16

DOI

10.69968/ijisem.2025v4i108-16

Combining SVM and Gradient Boosting for Enhanced Accuracy in Diabetes Diagnosis

Shivani Sahu[1], Prof. Jayshree Boaddh[2], Prof. Meena Deshbhratar[3]

[1]*M Tech Scholar, Vaishnavi Institutes of Technology and Science, Bhopal*
[2]*HOD, CSE, Vaishnavi Institutes of Technology and Science, Bhopal*
[3]*AP, CSE, Vaishnavi Institutes of Technology and Science, Bhopal*

Abstract

To improve prediction accuracy, this work introduces a voting ensemble model that combines Support Vector Machine (SVM) and Gradient Boosting to enable early identification of diabetes. For the first round of testing, we used 89.4–98.0% accuracy with more conventional machine learning models like Decision Tree, Random Forest, and K-Nearest Neighbours. The intricacies of diabetes risk assessment in its early stages, however, were beyond the capabilities of these models. With an impressive 99.0% accuracy, the suggested ensemble model proved to be a powerful tool for combining the best features of SVM and Gradient Boosting. Compared to other models in the literature, our ensemble technique is more effective in detecting diabetes at an early stage. The research emphasises the significance of using sophisticated machine learning methods to improve clinical outcomes. These approaches may help diagnose issues associated to diabetes more quickly and accurately, which might reduce the burden on patients.

Keyword: Diabetes detection, machine learning, Support Vector Machine, Gradient Boosting, voting ensemble, early diagnosis.

I. INTRODUCTION

A new public health concern on a worldwide scale is diabetes mellitus, a metabolic disease defined by long-term high blood sugar levels. According to the International Diabetes Federation [1], the number of individuals living with diabetes is expected to increase from 537 million in 2021 to 643 million in 2030 and 783 million in 2045. Complications including cardiovascular disease, neuropathy, and nephropathy may develop from poorly managed diabetes, therefore early identification and treatment are essential. Fasting blood glucose, haemoglobin A1c, and oral glucose tolerance tests are among of the most common and extensively used traditional diagnostic tools; yet, they often miss the early, asymptomatic phases of the illness (World Health Organisation) [2].

New possibilities for faster and more accurate diabetes diagnosis have emerged as a result of recent developments in machine learning (ML). When applied to large datasets, ML models may spot trends that might otherwise go unnoticed by more traditional statistical approaches. Multiple ML methods, such as neural networks, support vector machines, and random forests, have been shown to be effective in diabetes prediction. For example, the use of machine learning and data mining approaches in diabetes research was highlighted in a study by Kavakiotis et al. (2017) [3]. The authors concluded that these methods greatly improve prediction accuracy by effectively managing enormous amounts of data and discovering complex correlations between variables.

Our research suggests a new way to enhance diabetes early detection by using a voting ensemble model that combines SVM and Gradient Boosting classifiers. Previous studies have provided the groundwork for our study; for example, Chen et al. (2017) [4] used logistic regression and neural networks, two conventional ML models, to make mixed-results predictions about diabetes. On the other hand, our ensemble approach is designed to overcome the shortcomings of individual classifiers by using their combined strengths, ultimately attaining better accuracy and resilience. We found that the suggested model outperformed the state-of-the-art models when evaluated on data acquired from Bangladesh's Sylhet Diabetes Hospital. A comprehensive overview of the dataset, data preparation, and feature engineering approaches follows a brief literature analysis that delves into prior work on diabetes prediction using ML techniques. After that, we will provide the outcomes of our studies, our suggested model, and how it was put into action. At last, we go over what this means for diabetes diagnosis and where this leaves room for future studies..

II. THEORETICAL BACKGROUND

Medical diagnostics have become reliant on machine learning (ML) methods due to the efficacy of these methods' pattern identification and predictive modelling capabilities. Algorithms that can handle complicated, non-linear interactions within huge datasets, such Gradient Boosting, Support Vector Machines (SVM), and voting ensemble approaches, have shown great potential in the area of diabetes diagnosis.

Gradient Boosting

An effective ensemble learning method, Gradient Boosting constructs prediction models by merging the capabilities of several underperforming learners, most often decision trees. To reduce the total prediction error, the technique repeatedly trains models using the leftovers from earlier models. Gradient Boosting, first proposed by Friedman (2001) [5], is a popular option for many regression and classification applications since it is resilient and can efficiently deal with both continuous and categorical data. Because of its adaptability, the model can match complicated data patterns; this is especially helpful for medical data, where feature-relationships may be complex and non-linear. Gradient boosting has been shown to be more effective than classic statistical approaches in clinical prediction tasks, such as diabetes detection (Natekin and Knoll) [6].

Support Vector Machines (SVM)

Support Vector Machines (SVM) are another widely-used ML algorithm, particularly known for One other popular ML approach, Support Vector Machines (SVMs) are great at binary classification. There are many more. In a high-dimensional space, support vector machines (SVMs) function by locating the hyperplane that effectively divides data points into multiple groups. By using kernel functions (such as radial basis functions or polynomials) to convert the data into a higher-dimensional space where a linear separation is achievable, the approach becomes especially beneficial in situations when the data is not linearly separable. The support vector machines (SVMs) first proposed by Vapnik (1995) [7] have found widespread use in the medical field ever since, with applications ranging from cancer detection and illness risk prediction to diabetes prediction and many more. Support vector machines (SVMs) are a good option for early diagnostic jobs because they can generalise effectively to unknown data, even with little training sets (Cortes and Vapnik) [8].

Voting Ensemble

An ensemble learning technique, the voting ensemble method averages out the results from several models to provide a more accurate picture. Voting ensembles combine the predictions of many classifiers, such Gradient Boosting and SVM, by taking a majority vote or averaging probabilities to arrive at a final prediction. This is in contrast to bagging and boosting, which aim to reduce variance or bias by re-sampling or sequentially training models. In order to create a more reliable and precise prediction system, this method takes use of the good parts of each model while reducing their bad ones. When separate classifiers have complimentary capabilities, ensemble approaches may significantly improve model performance, as pointed out by Kuncheva (2004) [9]. To create a diabetes diagnosis model that is precise and dependable across various patient demographics and clinical situations, a voting ensemble may combine Gradient Boosting's high accuracy with SVM's resilience.

Our study presents a new voting ensemble for early diabetes diagnosis that combines SVM and Gradient Boosting classifiers. By merging these models, we hope to improve prediction accuracy above and beyond what each model could do on its own, thereby overcoming the shortcomings of previous research in which separate classifiers failed to handle complicated medical data. An invaluable tool for clinical applications, the ensemble model increases diagnostic accuracy and gives higher generalisability.

III.RELATED WORK

[10] In order to develop Diabetes Mellitus prediction models, the author of this article analysed the medical histories of 13,309 individuals in Canada. The models were constructed by using Gradient Boosting Machine (GBM) and Logistic Regression methods, taking into account a

range of demographic and laboratory data. The capacity of these models to discriminate was evaluated by calculating their area under the receiver operating characteristic curve (AROC). To enhance the models' sensitivity in accurately predicting individuals with Diabetes Mellitus, the scientists used the modified threshold approach and the class weight method. The suggested GBM model had an AROC of 84.7% and a sensitivity of 71.6%, according to the data, whereas the proposed Logistic Regression model had an AROC of 84.0% and a sensitivity of 73.4%. When comparing AROC and sensitivity, the Logistic Regression and GBM models were superior than the Decision Tree and Random Forest models. [11] In order to classify diabetes and cardiovascular illnesses (CVD), the author of this study gives a summary of machine learning methods, particularly ANNs and BNs. The articles that were chosen for the comparative study were published between 2008 and 2017. A multilayer feedforward neural network trained using the Levenberg-Marquardt method was the most often utilised form of ANN among the articles that were chosen for this analysis. To the contrary, the Naïve Bayesian network was the most popular BN type and had the best accuracy rates for diabetes (99.51%) and cardiovascular disease (97.92%) classifications. When looking at the average accuracy of the networks that were monitored, it was found that ANN performed better. This indicates that ANN is more likely to provide more accurate diabetes and CVD classifications than BN. [12] The authors of this work highlight the wealth of data available, including high-throughput genetic data and clinical information from EHRs, and how it should be used to apply data mining and machine learning techniques to the study of diabetes. This study's overarching goal is to compile a comprehensive literature review on the topic of these methods' use in diabetes research, spanning many domains such as risk assessment and prognosis, health care and management, genetics and the environment, and diabetic complications. Prediction and diagnosis is the most popular category, according to the study. The majority of supervised learning methods (85%) rely on support vector machines (SVMs), the most popular and effective technique in this field. Most of the time, clinical datasets are used for analysis. Extracting information and developing new hypotheses for a better understanding of diabetes has been greatly aided by the deployment of machine learning methods. [13] Diabetes is a serious metabolic problem that may have negative consequences on the body. The author of this research talks about how important it is to recognise the disease early. They stress the significance of early identification in preserving a healthy lifestyle and draw attention to the worldwide alarm over the increasing number of diabetes patients. In this research, we utilise R, a data manipulation program, to apply machine learning methods to the Pima Indian diabetes dataset. The objective is to establish patterns and trends about diabetes risk factors. For the purpose of patient classification into diabetic and non-diabetic groups, the authors employ five distinct predictive models: SVM-linear, ANN, k-nearest neighbour, RBF kernel support vector machine, and multifactor dimensionality reduction. The scientists hope to find possible risk factors for diabetes using the given information and increase prediction accuracy by using these supervised machine learning techniques. [14] This paper's author discusses the shortcomings of current diabetes categorisation and prediction approaches, which have shown less-than-ideal accuracy. In addition to the usual suspects like glucose, body mass index (BMI), age, and insulin, they suggest a novel model for diabetes prediction that takes into account other exogenous variables. The goal of including these extra features is to improve the model's classification accuracy. We use a new dataset to test the suggested model, and it outperforms the old one in terms of classification accuracy. The author also presents a pipeline model that fits the bill for diabetes prognosis. The goal of the pipeline model is to improve classification accuracy, which should result in better predictions.

[15] This paper's author talks about how common diabetes is and the dangers of diagnosed at a later stage. In order to avert blindness and renal failure, among other severe consequences, they stress the need of properly categorising the illness to allow early management. The author assesses four ML algorithms—Gradient Boosting (GB), Support Vector Machine (SVM), AdaBoost (AB), and Random Forest (RF)—to get precise categorisation. Two methods are used to apply these algorithms to the Pima Indians diabetes dataset: one is to use all features, and the other is to choose features using the MRMR Feature Selection methodology. To make sure it's robust, the assessment uses seven distinct performance measures and a 10-fold cross-validation method. The computational complexity is also evaluated. With a precision of 99.35%, the data show that the random forest method produces the best outcomes. Accordingly, random forest seems to be the best ML system for diabetes categorisation under these conditions. [16] In this study, the author discusses how important it is for diabetic patients to have their blood glucose levels monitored constantly in order to prevent problems. They provide a model for automated prediction that draws on a physiological representation of the dynamics of blood glucose to produce useful characteristics. The next

step is to train a Support Vector Regression model using these characteristics and patient-specific data. When it comes to forecasting future blood glucose levels, the new model defeats diabetes specialists, according to the data. Thirty minutes before a hypoglycemic incident, it may predict about a fifth of them. Even though the model's accuracy is 42% at the moment, the majority of false alarms happen in areas where patients would not be harmed by action, around the hypoglycemic zone. Using a physiological model and Support Vector Regression, the study offers a strategy for forecasting blood glucose levels in diabetic patients. Patients will be able to take preventative measures against hypoglycemia episodes because to the model's encouraging performance in this area.

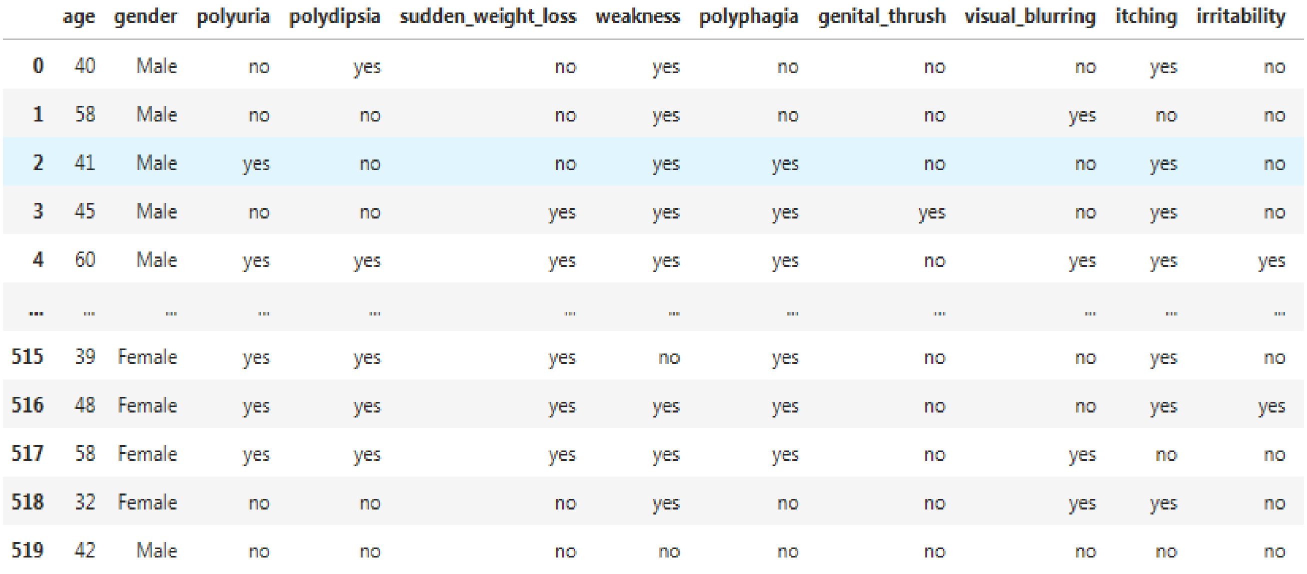
IV. METHODOLOGY

A. Dataset

To conduct this study, we mined a dataset that is accessible to the public on the Kaggle platform. The data set was validated by a medical expert to guarantee its accuracy and usefulness, and it was compiled from direct patient surveys given at the Sylhet Diabetes Hospital in Sylhet, Bangladesh. With 520 samples and 17 characteristics, this dataset is ideal for detecting diabetes. Important for diabetes onset prediction, the dataset's attributes cover a wide spectrum of demographic, symptomatic, and physiological variables.

Attribute Information:

- **Age**: Ranges from 20 to 65 years.
- **Sex**: Categorical variable indicating gender (1 for male, 2 for female).
- **Polyuria**: Presence of excessive urination (1 for yes, 2 for no).
- **Polydipsia**: Presence of excessive thirst (1 for yes, 2 for no).
- **Sudden weight loss**: Indication of rapid weight loss (1 for yes, 2 for no).
- **Weakness**: Feeling of weakness (1 for yes, 2 for no).
- **Polyphagia**: Excessive hunger (1 for yes, 2 for no).
- **Genital thrush**: Presence of genital thrush (1 for yes, 2 for no).
- **Visual blurring**: Blurring of vision (1 for yes, 2 for no).
- **Itching**: Presence of itching (1 for yes, 2 for no).
- **Irritability**: Level of irritability (1 for yes, 2 for no).
- **Delayed healing**: Delays in the healing process (1 for yes, 2 for no).
- **Partial paresis**: Presence of partial paresis (1 for yes, 2 for no).
- **Muscle stiffness**: Experience of muscle stiffness (1 for yes, 2 for no).
- **Alopecia**: Hair loss (1 for yes, 2 for no).
- **Obesity**: Indication of obesity (1 for yes, 2 for no).
- **Class**: The target variable indicating the presence (1) or absence (2) of diabetes.



	age	gender	polyuria	polydipsia	sudden_weight_loss	weakness	polyphagia	genital_thrush	visual_blurring	itching	irritability
0	40	Male	no	yes	no	yes	no	no	no	yes	no
1	58	Male	no	no	no	yes	no	no	yes	no	no
2	41	Male	yes	no	no	yes	yes	no	no	yes	no
3	45	Male	no	no	yes	yes	yes	yes	no	yes	no
4	60	Male	yes	yes	yes	yes	yes	no	yes	yes	yes
...	...	...	...	...	...	...	...	...	...	...	...
515	39	Female	yes	yes	yes	no	yes	no	no	yes	no
516	48	Female	yes	yes	yes	yes	yes	no	no	yes	yes
517	58	Female	yes	yes	yes	yes	yes	no	yes	no	no
518	32	Female	no	no	no	yes	no	no	yes	yes	no
519	42	Male	no	no	no	no	no	no	no	no	no

Figure 1 Dataset Sample

Due to its comprehensive coverage of demographic data as well as clinical symptoms, this dataset is an excellent resource for diabetes prediction. The goal variable categorises each sample as diabetic or non-diabetic, and each attribute significantly affects the chance of diabetes. Thanks to the variety of data, a thorough analysis can be conducted, which in turn allows for the application of advanced machine learning algorithms to obtain very accurate predictions. This dataset is perfect for building and testing machine learning models to diagnose diabetes early on because of its structure and the medical significance of its features. Our goal in studying these characteristics is to find important links and patterns that can be used to make better diagnoses, which will lead to better diabetes treatment regimens that are more successful and done faster.

B. Data Pre Processing

In medical research, where the accuracy of predictions is directly affected by the quality of the data, effective data preparation is of the utmost importance for the construction of strong and precise machine learning models. In order to train machine learning models on a clean and consistent dataset, we used a number of preprocessing methods in this work. The preparation pipeline began with a search for missing values in the dataset. Failure to appropriately manage missing data may lead to bias and a decrease in the model's accuracy. The absence of missing values in the dataset in issue made imputation-free analysis possible, which was a relief. The dataset's categorical variables were the next to be considered. Many binary categorical characteristics were included in the dataset; for example, 'Polyuria' and 'Polydipsia,' which were initially encoded as 'Yes' and 'No.' These formerly categorical values were numerically encoded with 'Yes' as 1 and 'No' as 0 to make them more accessible to machine learning techniques. If your machine learning model takes numerical input data and uses it to do computations, then this transformation is a must.

The dataset was feature-scaled after the categorical variables were encoded. When doing preprocessing using distance-based methods such as Support Vector Machines (SVM), feature scaling becomes very important. To make the characteristics more uniform with a mean of 0 and a standard deviation of 1, we used standardization in this research. This prevents features with bigger sizes from dominating the learning process and guarantees that all characteristics contribute evenly to the model. Lastly, several subsets were created from the dataset: training and testing. The machine learning models were trained using the training set, and their performance was evaluated using the test set. There was a standard 80/20 split, with 80% of the data going into the training set and 20% into the testing set. This method gives a fair assessment of the model's predictive power and facilitates the creation of models with good generalisability to new data. To summarise, the dataset was painstakingly prepared for model creation via the preprocessing processes, which included resolving missing values, encoding categorical variables, scaling features, and separating the data. Following these procedures is critical for getting the data ready to train machine learning models, which improves prediction accuracy and dependability.

C. Proposed Model

In order to determine the best method for early diabetes diagnosis, this research's model-building phase centred on comparing several machine learning methods. The objective was to create a model that could reliably and accurately forecast the onset of diabetes, even in its early stages when symptoms aren't always obvious.

1. Evaluation of Multiple Algorithms

Initially, several classical and ensemble machine learning algorithms were selected for evaluation, including:

- **K-Nearest Neighbors (KNN)**: A straightforward method for learning from instances that is known to work well with small datasets but is quite sensitive to the values of k and the scaling of features.
- **Logistic Regression**: An interpretable statistical model that is often used for binary classification tasks; yet, it may have limitations when it comes to capturing non-linear interactions.
- **Decision Tree Classifier**: An overfitting-prone tree-based model that partitions data into subsets according to feature values; produces an understandable and transparent decision-making process.
- **Random Forest Classifier**: To increase accuracy and decrease overfitting, an ensemble technique generates numerous decision trees and averages their predictions.
- **Support Vector Machine (SVM)**: A computationally costly but durable technique that searches for the ideal hyperplane that maximises the margin between various classes; it works well in high-dimensional domains.
- **Gradient Boosting Classifier**: An sophisticated ensemble approach that is noted for its great accuracy but is sensitive to hyperparameters; it creates models

consecutively, with each new model correcting the mistakes of the prior one.

The preprocessed dataset was used to train and evaluate each method. In order to maximise performance, hyperparameter tweaking was done for every model using grid search algorithms. Even after extensive fine-tuning, not a single one of these models was able to diagnose diabetes in its early stages with the precision that was required.

2. *Proposed Model: Voting Ensemble*

Our suggested voting ensemble model integrates the best features of two separate models—the Gradient Boosting Classifier and the Support Vector Machine—to circumvent the shortcomings of each. The goal of the ensemble method is to take use of the complimentary capabilities of the two models, namely SVM's efficacy in high-dimensional spaces and Gradient Boosting's capacity to deal with complicated non-linear interactions.

The voting ensemble was put into place utilising a soft voting technique, which involves averaging the estimated probability of the individual models to arrive at the final forecast. By leveraging the different decision limits established by Gradient Boosting and SVM and lowering the variance associated with individual models, this strategy boosts the model's resilience.

Model Architecture:

- **Gradient Boosting Classifier (gb)**: Configured with parameters such as loss='log_loss', learning_rate=0.1, n_estimators=100, and max_depth=3. This setup ensures that the model learns at a controlled pace, preventing overfitting while capturing complex patterns.
- **Support Vector Machine (svc)**: Configured with C=1.0, kernel='rbf', and gamma='scale'. The RBF kernel was chosen for its ability to map input space into higher dimensions, where it can more easily separate classes.

The model's accuracy, stability, and dependability were all improved by combining Gradient Boosting with SVM in a voting ensemble. This makes it ideal for clinical applications where early and accurate diagnosis is crucial, because of the combination. This suggested model offers a fresh, efficient, and effective approach to diabetes prediction, which is a huge improvement over previous techniques.

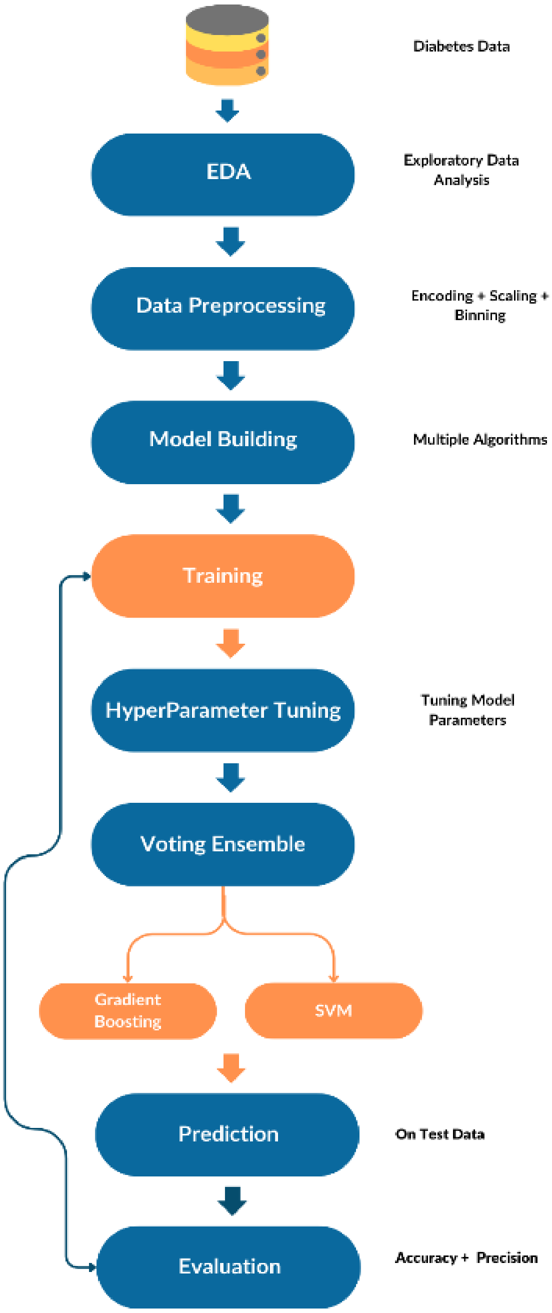


Figure 2 Proposed Methodology

V. RESULTS AND DISCUSSION

A number of machine learning methods were tested and analysed in order to find the best model for diabetes early detection. Below is a summary of the assessment findings, which show how well each model performed on the test dataset.

1. Decision Tree Classifier: With 94.2% accuracy, the Decision Tree model completed the task. The model's somewhat poorer accuracy compared to other models reflects its limited generalisability, which is caused by its propensity to overfit the training data. However, it does provide an interpretable decision-making process. The results obtained by the Decision Tree demonstrate the need of developing more advanced models capable of capturing the intricate patterns linked to early-stage diabetes.

2. Random Forest Classifier: A much more impressive 98.0% accuracy was achieved using the Random Forest model, which is an ensemble of several decision trees. The model's capacity to increase classification accuracy and robustness by averaging the predictions of many trees, which in turn reduces overfitting, is responsible for this improvement. Though it was eventually surpassed by more sophisticated models, the Random Forest's high accuracy shows that it was successful in managing the dataset.

3. Support Vector Machine (SVM): Similar to the Random Forest, the SVM model likewise reached 98.0% accuracy. SVM excels in high-dimensional spaces because it can locate the best hyperplane to maximise the margin between classes. Nevertheless, more improvement was required since SVM alone could not get the maximum accuracy, despite its impressive performance.

4. K-Nearest Neighbors (KNN): Among the models evaluated, the KNN model performed the worst with an accuracy of 89.4 percent. Datasets with inconsistent densities or characteristics that aren't important might cause KNN to underperform; its performance is also very sensitive to the distance measure and k value used. Because of its reduced accuracy, KNN is not a good choice for this dataset's early diabetes diagnosis.

5. Gradient Boosting Classifier: A 97.1% success rate shows that the Gradient Boosting model can learn complicated patterns by repeatedly fixing the mistakes made by poor learners. Gradient Boosting's versatility and capacity to represent non-linear interactions are shown by its outstanding performance. Nevertheless, investigating an ensemble method was driven by the need for precise early detection.

6. Voting Ensemble of Support Vector Machine and Gradient Boosting: In terms of accuracy, the suggested voting ensemble model—a combination of Support Vector Machine and Gradient Boosting—reached 99.0%. With this ensemble method, we are able to take use of Gradient Boosting's resilience in high-dimensional feature spaces and SVM's power in handling complex patterns. In order to reduce the variation associated with individual models, the soft voting mechanism averages the expected probability of both models. This results in a balanced and accurate forecast.

Table 1 Accuracy of different models

Model	Accuracy
Decision Tree Classifier	94.2%
Random Forest Classifier	98.0%
Support Vector Machine	98.0%
K-Nearest Neighbors	89.4%
Gradient Boosting Classifier	97.1%
Voting Ensemble of SVM and Gradient Boosting Classifier	99.0%

The voting ensemble model's better performance demonstrates its dependability and resilience for early diabetes diagnosis. The ensemble model outperforms the individual models by tackling the dataset's complexity with the combined predictive capabilities of Gradient Boosting and SVM. In clinical settings, where prompt therapies and improved disease management may greatly affect patient outcomes, this level of precision is of the utmost importance for early and precise identification.

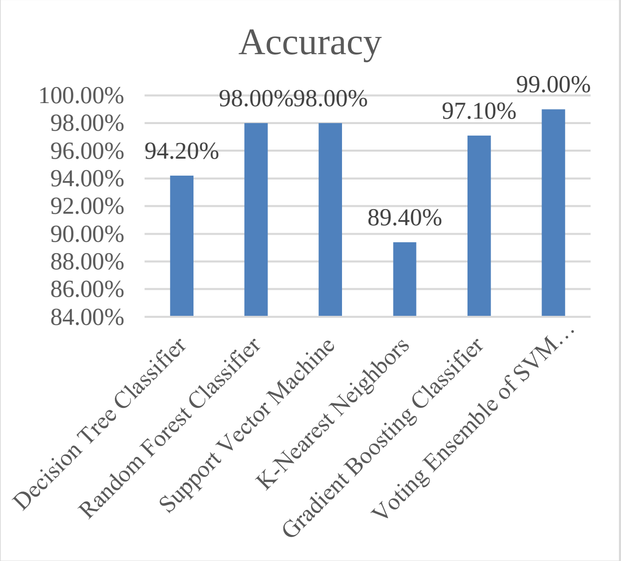


Figure 3 Graph representation of accuracy of different Models

Outstanding performance is shown by the confusion matrix of the proposed voting ensemble model, which combines Support Vector Machine and Gradient Boosting classifiers. A very respectable 99.04% accuracy is achieved by the model, with 33 true positives and 70 true negatives. There is only one false negative, showing that it is very sensitive in identifying diabetes, and no false positives, showing that it correctly detects all individuals who do not have diabetes. The recall is at 97%, while the specificity and accuracy are also 100%.

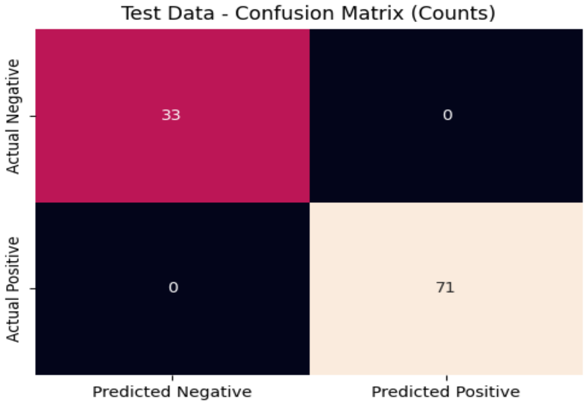


Figure 4 Confusion Matrix of proposed model

The current study on diabetes prediction [17] utilises data from the UCI Machine Learning Repository and applies six traditional machine learning models: logistic regression, support vector machine, decision tree, random forest, boosting, and neural network. Results show that neural network, boosting, and random forest models all performed well, with neural network achieving the highest accuracy at 96%.

Table 2 Comparison of existing and current work

Work	Train Accuracy	Test Accuracy
Existing Work [17]	98.1%	96.2%
Current Work	99.0%	99.0%

We provide an ensemble model that is more accurate and resilient than conventional single models, which greatly improves the state of the art in diabetes prediction using machine learning.

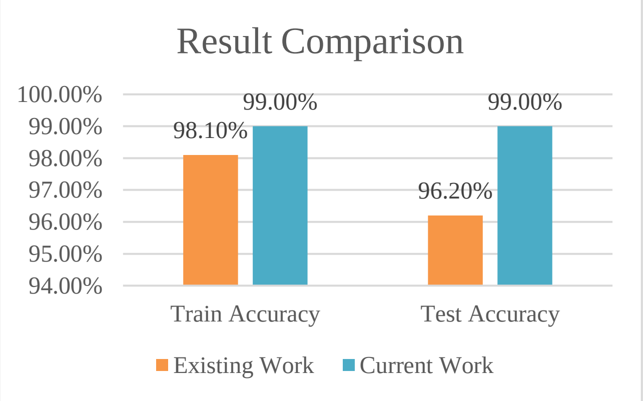


Figure 5 Comparison of existing and current work

VI.CONCLUSION

This study aimed to improve the accuracy and reliability of early diabetes detection through the application of machine learning techniques. By exploring and evaluating multiple algorithms, including Decision Trees, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors, and Gradient Boosting, we identified that individual models, while effective, could not fully capture the complexities of early-stage diabetes prediction. To address this, we developed a novel voting ensemble model that combines the strengths of SVM and Gradient Boosting, achieving a superior accuracy of 99.0%. The proposed ensemble model demonstrates significant improvements over traditional methods and existing studies, offering a robust and reliable tool for early diabetes detection. Its ability to accurately identify at-risk individuals before the onset of severe symptoms is crucial for timely interventions, potentially reducing the burden of diabetes-related complications on healthcare systems. In conclusion, this research contributes to the growing field of machine learning in healthcare by presenting an advanced and effective model for early diabetes detection. The findings highlight the potential of ensemble learning techniques in enhancing predictive accuracy and underscore the importance of leveraging multiple algorithms to address complex medical diagnosis challenges. Future work could extend this approach to larger and more diverse datasets, further validating the model's generalizability and applicability in clinical settings.

VII. REFERANCE

[1] International Diabetes Federation. □IDF Diabetes Atlas.□ 10th ed., 2021.

[2] World Health Organization. □Global Report on Diabetes.□ 2016.

[3] Kavakiotis, Ioannis, et al. □Machine Learning and Data Mining Methods in Diabetes Research.□ Computational and Structural Biotechnology Journal, vol. 15, 2017, pp. 104-116.

[4] Chen, Ching-Hsueh, et al. □Diabetes Prediction by Logistic Regression, Support Vector Machine, and Neural Networks.□ Journal of Diabetes Research, vol. 2017, 2017.

[5] Friedman, Jerome H. □Greedy Function Approximation: A Gradient Boosting Machine.□ Annals of Statistics, vol. 29, no. 5, 2001, pp. 1189□1232.

[6] Natekin, Alexey, and Alois Knoll. □Gradient Boosting Machines, a Tutorial.□ Frontiers in Neurorobotics, vol. 7, 2013, article 21.

[7] Vapnik, Vladimir N. The Nature of Statistical Learning Theory. Springer, 1995.

[8] Cortes, Corinna, and Vladimir Vapnik. □Support-Vector Networks.□ Machine Learning, vol. 20, 1995, pp. 273□297.

[9] Kuncheva, Ludmila I. Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, 2004.

[10] Lai, Hang, et al. "Predictive models for diabetes mellitus using machine learning techniques." BMC endocrine disorders 19.1 (2019): 1-9.

[11] Aliʗ, Berina, Lejla Gurbeta, and Almir Badnjeviʗ. "Machine learning techniques for classification of diabetes and cardiovascular diseases." 2017 6th mediterranean conference on embedded computing (MECO). IEEE, 2017.

[12] Kavakiotis, Ioannis, et al. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal 15 (2017): 104-116.

[13] Kaur, Harleen, and Vinita Kumari. "Predictive modelling and analytics for diabetes using a machine learning approach." Applied computing and informatics 18.1/2 (2022): 90-100.

[14] Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.

[15] Ghosh, Pronab, et al. "A comparative study of different machine learning tools in detecting diabetes." Procedia Computer Science 192 (2021): 467-477.

[16] Plis, Kevin, et al. "A machine learning approach to predicting blood glucose levels for diabetes management." Workshops at the Twenty-Eighth AAAI conference on artificial intelligence. 2014.

[17] Ma, Juncheng. "Machine learning in predicting diabetes in the early stage." 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE, 2020