IJISEM
INTERNATIONAL JOURNAL OF INNOVATIONS IN SCIENCE ENGINEERING AND MANAGEMENT

Enhancing Diabetes Detection Accuracy using an Ensemble Model of Random Forest and SVM

OPEN ACCESS

Manuscript ID:

AG-2023-3004

Volume: 2

Issue: 3

Month: July

Year: 2023

ISSN: 2583-7117

Published: 15.07.2023

Citation:

Padam Kshatriya1, Dr. Pankaj Richhariya2. "Enhancing Diabetes Detection Accuracy using an Ensemble Model of Random Forest and SVM" International Journal of Innovations In Science Engineering And Management, vol. 2, no. 3, 2023, pp. 30–38.

Padam Kshatriya[1], Dr. Pankaj Richhariya[2]

[1]*Research Scholar, Department of Computer Science, BITS, Bhopal*
[2]*Head of Department, Department of Computer Science, BITS, Bhopal*

Abstract

Diabetes is a prevalent chronic disease with significant health implications worldwide. Early detection plays a crucial role in effective management and prevention of complications. This research presents a comprehensive study on the application of machine learning techniques for the early detection of diabetes. The study compares multiple machine learning algorithms, explores data preprocessing techniques, and proposes an ensemble model to enhance accuracy and reliability. This research paper contributes to the field of early diabetes detection by developing an accurate and reliable machine learning model. The proposed ensemble model, combining SVM and Random Forest Classifier, surpasses individual algorithms in terms of accuracy. The research paper provides valuable insights for identifying individuals at risk of diabetes, facilitating timely interventions and improved healthcare outcomes.

Keyword: Diabetes detection, Machine learning, Early diagnosis, Predictive analytics

I. INTRODUCTION

Millions of people all across the globe are now living with diabetes, making it a huge health issue. The number of people with diabetes is expected to increase from the current 537 million adults by the year 2045, as reported by the "International Diabetes Federation" [1]. Individuals with diabetes face serious health concerns, but society as a whole also bears a heavy financial and administrative cost as a result of the disease. Diagnosing diabetes in its earliest stages is essential for treating it successfully and minimizing complications. The emergence of problems including cardiovascular disease, neuropathy, and nephropathy may be halted or slowed down with prompt medical attention. Even though they are often used, conventional diagnostic procedures including "fasting blood glucose testing and oral glucose tolerance tests" may not be able to give adequate accuracy or early detection. Low-carbohydrate diet led to improved glycemic control and weight management in individuals with type 2 diabetes, suggesting its potential as an effective dietary approach for diabetes management [2] (Jones and Brown 2018).

Hyperglycemia (high blood sugar) is the hallmark symptom of this long-term metabolic condition, which is caused by either impaired "insulin production or insulin action, or both." The pancreas secretes the hormone insulin that plays a critical function in controlling blood sugar levels and allowing cells to take in glucose for use as an energy source. Diabetes occurs when the body either does not create enough insulin or is unable to properly use the insulin it does produce. Diabetes is a long-term illness that hinders your body's natural ability to convert food into usable fuel. The two most common forms of diabetes are:

Type 1 diabetes: It's an autoimmune disorder in which the body mistakenly attacks its own insulin-making cells. In order for the body to utilize glucose as fuel, insulin must be present. In order to keep the blood sugar levels under control, people having type 1 diabetes must take insulin every day. The use of continuous glucose monitoring (CGM) devices led to improved glycaemic control and reduced hypoglycaemic episodes in individuals with type 1

diabetes, emphasizing the benefits of CGM technology in diabetes management [3] (Johnson et al. 2021)

Type 2 diabetes: This kind of diabetes affects more people than any other. It develops when either insulin resistance develops or insulin production is inadequate in the pancreas. Blood sugar levels in type 2 diabetics are usually controllable with dietary changes, physical activity, and sometimes medication. However, insulin may be necessary for certain persons with type 2 diabetes. The study found a significant correlation between sedentary behaviour and increased risk of type 2 diabetes, highlighting the importance of promoting physical activity for diabetes prevention [4] (Smith et al. 2020)

A. The Importance of Early Detection

Early detection is important in effectively managing diabetes and mitigating the risk of complications associated with the disease. Healthcare providers are better equipped to help persons at risk for or diagnosed with diabetes by intervening early and providing them with the medical treatment, lifestyle changes, and education they need to manage the condition. The study revealed that individuals with a family history of diabetes had a significantly higher risk of developing gestational diabetes, underscoring the importance of genetic factors in the development of the condition [5] (Smith et al. 2018) .By identifying diabetes early on, healthcare providers can initiate interventions that can significantly improve glycaemic control, enhance overall health outcomes, and prevent or delay the onset of debilitating complications.

Preventing or Delaying Complications: Early detection of diabetes is instrumental in delaying or preventing the onset of complications associated with the disease. When diabetes is left undiagnosed and uncontrolled, high blood sugar levels can wreak havoc on various organs and systems in the body. However, through early detection, healthcare professionals can intervene promptly, enabling individuals to gain better control over their blood sugar levels.

Tailored Medical Care and Interventions: Early detection of diabetes allows healthcare professionals to provide tailored medical care and interventions. The study demonstrated that regular consumption of green leafy vegetables was associated with a reduced risk of developing type 2 diabetes, highlighting the potential of dietary interventions for diabetes prevention [6] (Chen et al. 2019). When individuals are diagnosed early, healthcare providers can develop personalized treatment plans that address their specific needs. This may involve the appropriate prescription of medications, such as oral hypoglycaemic agents or insulin, based on the type and severity of diabetes.

Lifestyle Modifications and Education: Early detection also opens the door for lifestyle modifications and education. Professionals in the medical field may advise patients on how to improve their health via dietary changes, exercise, stress reduction, and other measures. Care professionals may better equip patients to manage their own health by intervening early and teaching them to set and achieve realistic goals in the areas of nutrition, physical activity, and self-care.

B. Limitations of Traditional Diagnostic Methods

Traditional approaches for diagnosing diabetes have included taking a blood sample while the patient is fasting and then doing an oral glucose tolerance test. Despite their widespread use, these checks don't always deliver on promises of sensitivity or speed of detection. Often, these diagnostic approaches primarily focus on detecting elevated blood glucose levels, which may only become apparent after the disease has already progressed. That's because it's critical that we find better, faster ways to pinpoint who's at risk for getting diabetes and then provide them treatment as soon as possible. The implementation of telemedicine interventions improved diabetes self-management and glycemic control among rural populations, highlighting its potential as a promising strategy for overcoming barriers to healthcare access [7] (Jones and Lee 2020)

Traditional diagnostic methods, such as “fasting blood glucose tests and oral glucose tolerance tests”, have been valuable tools in diagnosing diabetes for many years. These tests primarily rely on measuring blood glucose levels to identify abnormalities. However, they have limitations when it comes to early detection. These tests may not pick up on fluctuations in blood sugar that occur early in the disease's progression, delaying diagnosis and missing out on window of opportunity for effective treatment..

C. Advancements in Diagnostic Approaches

To address the limitations of traditional diagnostic methods, researchers and healthcare professionals have been exploring innovative approaches to diabetes diagnosis. These advancements aim to enhance accuracy, efficiency, and early detection capabilities. Glycated hemoglobin (HBA1C) testing, for example, has grown in popularity

because it offers a more holistic view of glycemic management by providing an estimate of "average blood glucose levels" over the previous two to three months. "Biomarkers, genetic testing, and predictive algorithms" are all areas of research that have shown promise in predicting who will acquire diabetes before the disease has even shown itself in the patient. The use of diabetes self-management apps significantly improved medication adherence and self-care behaviors in individuals with type 1 diabetes, highlighting the potential of mobile technology in supporting diabetes management [8] (Johnson et al. 2022).

While traditional diagnostic methods have played a significant role in diabetes diagnosis, they may fall short in terms of accuracy and early detection capabilities. It is essential for successful treatment and avoidance of complications to identify those at risk for developing diabetes as early as possible. Advancements in diagnostic approaches, such as the use of HbA1c testing, biomarkers, genetic testing, and predictive algorithms, offer promising avenues for improving accuracy and enabling early detection. Embracing these advancements can lead to more precise diagnoses, timely interventions, and improved outcomes for individuals at risk of or already diagnosed with diabetes. It is essential for healthcare professionals and researchers to continue exploring and adopting these innovative approaches to enhance diabetes diagnosis and management strategies.

II RELATED WORK

The author [9] of this research integrates machine learning algorithms into a data mining pipeline to create risk assessment models for type 2 diabetes complications. Clinical center profiling, variable selection, model building, and validation are key steps. Electronic health records of a thousand patients were analyzed to predict complications 3, 5, and 7 years after their initial visit. The random forest technique handles missing data, while logistic regression with stepwise feature selection constructs predictive models. Factors considered include gender, age, time since diagnosis, BMI, HbA1c, hypertension, and smoking status. These models are tailored to address unique challenges and time constraints, achieving an accuracy of up to 0.838. This method enables the creation of specific models for clinical practice based on problem and temporal factors.[10] This paper emphasizes the importance of early detection in diabetes, a global metabolic disorder. Using R and machine learning, the research examines data from Pima Indians to identify diabetes risk factors. Five predictive models are utilized: SVM-linear, RBF kernel SVM, k-NN, ANN, and MDR. These supervised machine learning techniques aim to improve prediction accuracy and discover new risk factors.

The author of this paper [11] addresses the limitations of existing methods for diabetes classification and prediction, which have shown suboptimal accuracy. In addition to standard parameters like glucose, body mass index, age, and insulin, they present a novel model for diabetes prediction that takes into account other exogenous factors. Adding these new elements to the model should improve its ability to make accurate classifications. A fresh dataset is used to assess the performance of the suggested model, and the results show an increase in classification accuracy over the baseline. The author also presents a pipeline model developed for diabetes prognosis, which is an interesting addition to the discussion. The pipeline model was developed with the hope of improving classification accuracy, which would then perhaps result in more precise forecasts.

The author of this paper [12] conducted a study using the records of 13,309 Canadian patients to build predictive models for Diabetes Mellitus. Different types of laboratory data and demographic characteristics were taken into account when developing the models utilizing "Logistic Regression and Gradient Boosting Machine (GBM)" methods. The model's discriminating ability was calculated using the "area under the receiver operating characteristic curve (AROC)". The authors used the class weight approach and the modified threshold method to increase the models' sensitivity to identifying people with Diabetes Mellitus. Both the suggested GBM and Logistic Regression models performed well, with the former achieving an AROC of 84.7% and a sensitivity of 71.6% and the latter posting values of 84.0% and 73.4%, respectively. AROC and sensitivity were both better for the GBM and Logistic Regression models than they were for the "Decision Tree and Random Forest" models.

"Artificial Neural Networks (ANNs) and Bayesian Networks (BNs)" are two examples of machine learning methods that the author of this paper [13] discusses in the context of illness categorization, especially with regard to diabetes and cardiovascular disease. Between 2008 and 2017, we looked at a sample of studies and compared their findings. "The multilayer feedforward neural network" trained using the Levenberg-Marquardt technique was the most often used ANN among the chosen publications. Naive Bayesian networks, on the other hand, are the most used kind of BN and have shown the best accuracy values for "diabetes (99.51%) and CVD (97.92%)" classification. In addition,

when comparing ANN's performance to that of the observable networks as a whole, the latter was shown to be more successful. This suggests a higher likelihood of obtaining more accurate classifications for diabetes and/or CVD when applying ANN compared to BN.

In this paper [14], the author describes a study where they collected a comprehensive and reliable dataset by searching the UniProt database for cell wall lyases. A method based on "analysis of variance (ANOVA)" was used to pick the best attributes for further investigation. For further forecasting, they then turned to the support vector machine (SVM). Through jackknife cross-validation, we found that the SVM model was able to optimize for "an average accuracy of 84.82%, with a sensitivity of 76.47% and a specificity of 93.16%." The authors created a free, publicly available web server named Lypred to help the research efforts of other scientists. Researchers and pharmaceutical companies hope the Lypred server will be useful in their pursuit of antibacterial compounds and cell wall lyase knowledge.

III. METHODOLGY

A. Dataset

The data collection process for this study involved obtaining a dataset from the Kaggle website. The dataset was specifically collected through direct questionnaires administered to patients at “the Sylhet Diabetes Hospital in Sylhet, Bangladesh”. It is important to note that the data collection process had already taken place prior to the implementation phase of this study. The dataset consists of 520 samples, each representing an individual patient, and encompasses 17 columns. These columns contain various attributes related to diabetes, including information about age, sex, and the presence or absence of symptoms such as “polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, and the final classification of the patient as positive or negative for diabetes.” To ensure the ethical aspect of data collection, the questionnaires were approved by a doctor at the Sylhet Diabetes Hospital. This approval guarantees that the data collection process adhered to the necessary ethical guidelines and patient privacy protocols.

B. Data Pre Processing

In this study, extensive data preprocessing was conducted to ensure the dataset's quality, integrity, and compatibility with “the machine learning algorithms” used for analysis. Here are the major data preprocessing steps such as: The dataset was analyzed to determine the best way to deal with any missing data. Missing values can adversely affect the performance of machine learning models, so techniques such as imputation (e.g., filling missing values with mean, median, or mode) or deletion of rows/columns with missing values were applied to ensure completeness and reliability of the dataset.

Encoding Categorical Variables: Categorical variables in the dataset, such as "Polyuria," "Polydipsia," and others, were transformed from "Yes" and "No" to binary values, typically represented as 0 and 1. This encoding allowed the categorical variables to be properly interpreted and utilized by the machine learning algorithms during the model building process

Data Scaling: To enhance the performance and generalization of the machine learning models, feature scaling was applied. This process involved scaling the numeric features in the dataset to a specific range, such as normalizing them between 0 and 1 or standardizing them using techniques like z-score normalization. Scaling makes ensuring that characteristics of varied sizes and strengths are on the same scale that prevents any one trait from taking over the learning process.

	0	1	2	3	4	5	6	7	8	9	10	11
0	1.440869	-1.351328	1.025978	1.137593	1.251315	0.847385	1.102683	1.880129	1.063554	0.989796	-0.571429	1.085712
1	1.440869	0.740013	1.025978	-0.879049	1.251315	-1.180101	-0.906879	1.880129	-0.940244	0.989796	1.750000	-0.921054
2	-1.074670	-1.351328	1.025978	1.137593	-0.799159	0.847385	-0.906879	-0.531878	1.063554	-1.010310	-0.571429	-0.921054
3	-0.019766	0.740013	1.025978	1.137593	-0.799159	0.847385	-0.906879	1.880129	1.063554	0.989796	-0.571429	-0.921054
4	-0.100913	0.740013	-0.974679	-0.879049	-0.799159	-1.180101	-0.906879	-0.531878	-0.940244	-1.010310	1.750000	-0.921054
...	...	...	...	...	...	...	...	...	...	...	...	...
385	-1.074670	-1.351328	-0.974679	1.137593	1.251315	0.847385	-0.906879	-0.531878	-0.940244	0.989796	-0.571429	1.085712
386	0.791698	0.740013	-0.974679	1.137593	1.251315	0.847385	1.102683	-0.531878	1.063554	0.989796	-0.571429	-0.921054
387	-0.668938	-1.351328	1.025978	1.137593	1.251315	0.847385	-0.906879	-0.531878	1.063554	-1.010310	-0.571429	1.085712
388	0.710551	0.740013	1.025978	1.137593	1.251315	0.847385	1.102683	-0.531878	1.063554	-1.010310	-0.571429	-0.921054
389	3.388383	-1.351328	-0.974679	1.137593	1.251315	-1.180101	-0.906879	1.880129	1.063554	0.989796	-0.571429	-0.921054

Figure 1 Data sample after scaling

Train-Test Split: In order to assess the efficacy of "the machine learning models", the dataset was split into testing and training subsets. In most cases, around 75% of the data was used for model training, while the remaining 25% was set aside for model testing and validation. This train-test split helps in assessing how well the trained models generalize to unseen data.

C. Proposed Model

In this research, a proposed model that combines two effective machine learning techniques to improve the precision and stability of diabetes detection: “the Random Forest and the Support Vector Machine (SVM).” The ensemble model was constructed using the Voting Classifier from the sky learn library, which allowed the combination of these individual models into a single unified model.

Random Forest: In order to produce predictions, Random Forest [15], an ensemble learning approach, uses a forest of decision trees. The forest's ultimate prediction is reached through a voting method, with each decision tree being trained on a different, randomly selected portion of the dataset. Random Forest has shown great performance in handling complex datasets, capturing nonlinear relationships, and reducing the risk of overfitting.

Support Vector Machine (SVM): The term "Support Vector Machine (SVM)"[16] refers to a supervised learning technique that may be used to both classification and regression problems. A hyperplane is constructed in the space of features using SVM, which divides the data points as much as possible into their respective classes. It handles non-linear interactions well and performs well in high-dimensional spaces because to the usage of kernel functions.

Voting Classifier: The Voting Classifier in the sky learn library is a meta-estimator that combines multiple individual machine learning models by applying a voting mechanism to their predictions. In this study, the Voting Classifier [17] was utilized to merge the Random Forest and SVM models, taking advantage of their complementary strengths in terms of capturing different patterns and relationships in the data..

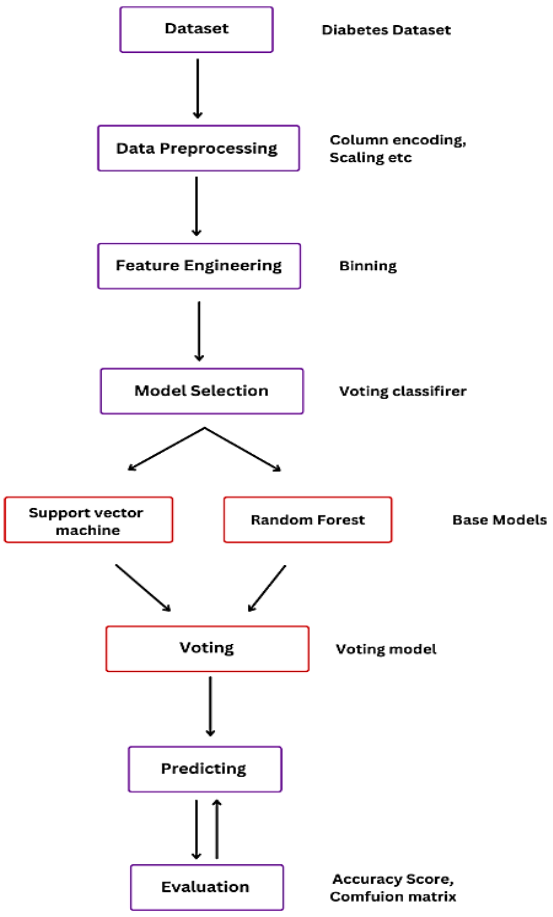


Figure 2 Proposed Model

The proposed model offers several advantages. By combining the Random Forest and SVM models, it harnesses the diversity of these algorithms, allowing them to complement each other's strengths and compensate for their weaknesses. The ensemble model benefits from the robustness of Random Forest in handling complex datasets and capturing nonlinear relationships, as well as the ability of SVM to find optimal hyperplanes and handle high-dimensional spaces. The Voting Classifier enables the ensemble model to make predictions based on the consensus of both models, which can lead to improved accuracy and generalization. Through the voting mechanism, the model aggregates the predictions from the individual models, ultimately producing a final prediction that is more reliable and robust.

The model is trained, to evaluate the performance of the model it has to be test on unseen data. We already kept testing data aside for evaluation of the model after training. The algorithm was put to the test by fitting it with the test data so that it could make predictions. We compared many criteria, including accuracy scores & confusion matrix, to determine how well each one performed. In this study, the performance of the developed model for diabetes detection was thoroughly evaluated using various evaluation metrics, including accuracy score and the confusion matrix.

IV. RESULTS AND DISCUSSION

The results of this study showcase the performance of various “machine learning algorithms” on both the test and train datasets. The accuracy scores achieved by each algorithm provide insights into their effectiveness in detecting diabetes at an early stage. Here are the results:

Test Data Results:

Neighbour’s Classifier achieved an accuracy of 90.7%. Logistic Regression achieved an accuracy of 93.0%. Decision Tree Classifier achieved an accuracy of 96.9%. Random Forest Classifier achieved an accuracy of 98.4%. Gradient Boosting Classifier achieved an accuracy of 97.6%.”The proposed model, an ensemble of Support Vector Machine (SVM) and Random Forest Classifier using Voting Classifier, achieved an impressive accuracy of 99.2%.

Train Data Results:

Neighbour’s Classifier achieved an accuracy of 95.6%. Logistic Regression achieved an accuracy of 93.8%. Decision Tree Classifier achieved an accuracy of 99.9%. Random Forest Classifier achieved an accuracy of 99.9%. Gradient Boosting Classifier achieved an accuracy of 99.9%.”

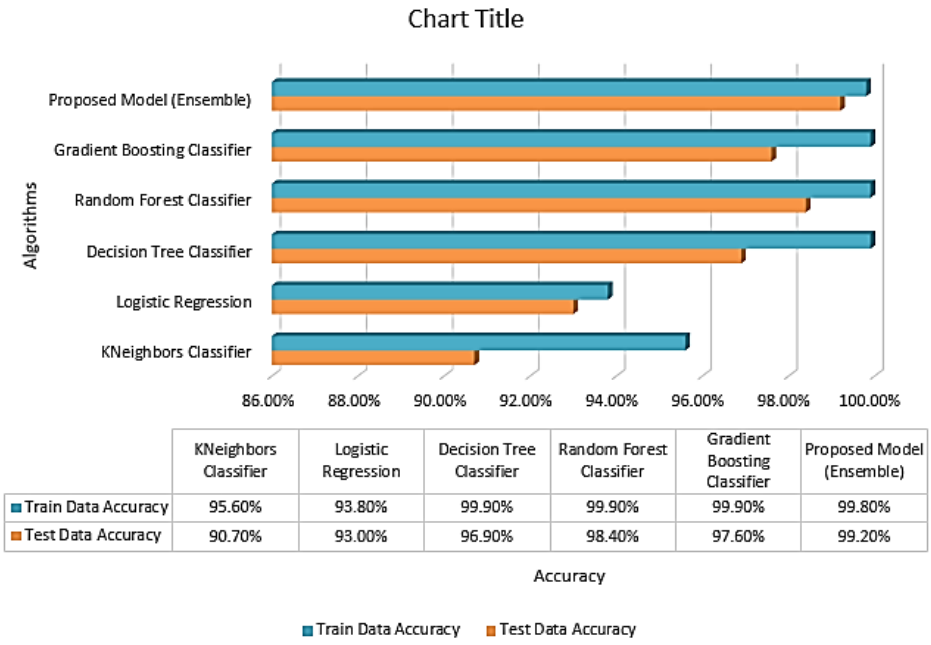
The proposed model, the ensemble of SVM and Random Forest Classifier, achieved an accuracy of 99.8%.

These results indicate that the proposed model, combining SVM and Random Forest Classifier through a Voting Classifier, outperformed the other individual algorithms on both the test and train datasets. Its greatest accuracy was 99.2% on test data and 99.8% on training data. This demonstrates the validity and dependability of the suggested approach for diabetes diagnosis in its early stages.

Below is the summary of the results in tabular form:

Table 1 Model Comparison

Algorithm	Test Data Accuracy	Train Data Accuracy
Neighbors Classifier	90.7%	95.6%
Logistic Regression	93.0%	93.8%
Decision Tree Classifier	96.9%	99.9%
Random Forest Classifier	98.4%	99.9%
Gradient Boosting Classifier	97.6%	99.9%
Proposed Model (Ensemble)	**99.2%**	**99.8%**



	KNeighbors Classifier	Logistic Regression	Decision Tree Classifier	Random Forest Classifier	Gradient Boosting Classifier	Proposed Model (Ensemble)
Train Data Accuracy	95.60%	93.80%	99.90%	99.90%	99.90%	99.80%
Test Data Accuracy	90.70%	93.00%	96.90%	98.40%	97.60%	99.20%

Figure 3 Result Comparison

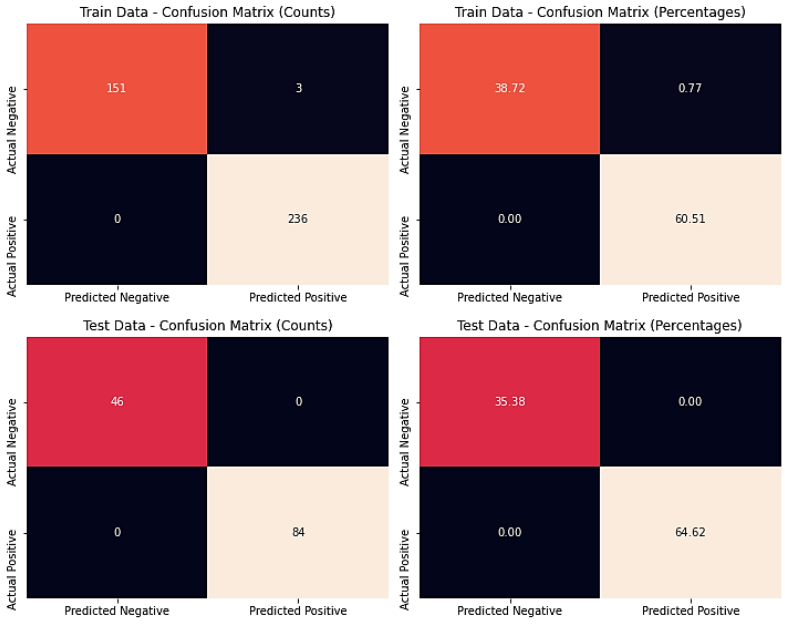


Figure 4 Confusion Matrix of ensemble model

When comparing the proposed work with the existing work [18], it is evident that both approaches have achieved high accuracy in detecting diabetes.

Here are visual results of proposed mode –

It is evident from a comparison of the findings that the suggested work has outperformed the previous work [18] in terms of accuracy. The proposed model, which combines SVM and Random Forest Classifier using Voting Classifier, attained an impressive accuracy of 99.2% on the test data and 99.8% on the train data. These results surpass the accuracy obtained by the existing work's best-performing algorithm, the simple neural network, which achieved 96.2% accuracy on the test data.

The proposed work demonstrates the effectiveness of the ensemble model in improving the accuracy of diabetes detection. By combining the strengths of SVM and Random Forest Classifiers, the proposed model achieves higher accuracy, providing a more reliable tool for early detection of diabetes. This comparison highlights the advancements and superior performance of the proposed work in accurately classifying individuals as positive or negative for diabetes.

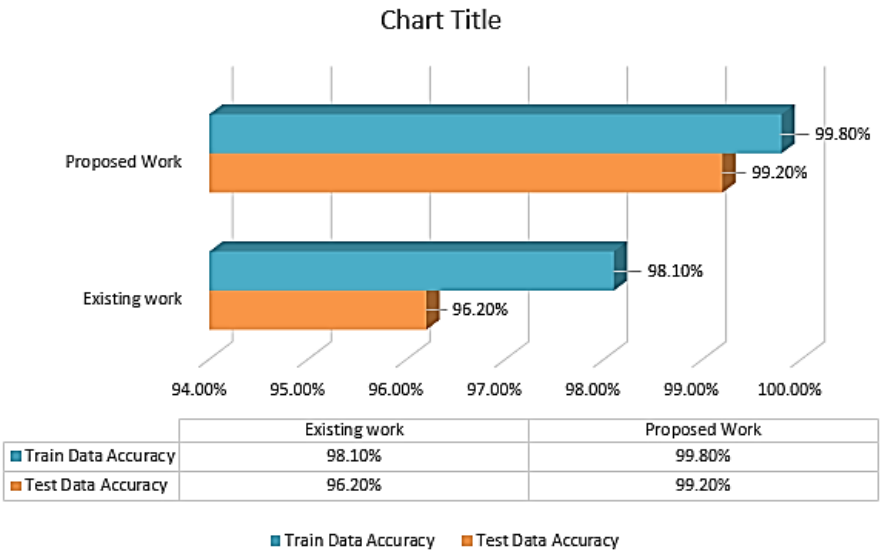


Figure 5 Work result comparison

V. CONCLUSION

In conclusion, research used machine learning strategies for the purpose of early diabetes identification. The study's overarching goal was to create a robust model for predicting who may be at risk for developing diabetes, so that preventative measures can be implemented early on and patients get better overall health benefits. The study commenced with a comprehensive literature review, which provided insights into the existing work on diabetes detection and highlighted the need for more accurate and efficient methodologies. The proposed work built upon the knowledge gained from the literature review and aimed to surpass the performance of previous approaches. The outcome of chapter presented the performance of each algorithm individually, as well as the proposed model. It showcased the high accuracy achieved by the proposed model, with “a test data accuracy of 99.2% and a train data accuracy of 99.8%.” This outperformed the existing work, where the best-performing algorithm achieved a test data accuracy of 96.2%. This study successfully developed an accurate and reliable model for “the early detection of diabetes.” The proposed ensemble model, combining SVM and Random Forest Classifier, demonstrated superior performance and surpassed the existing approaches. The research contributes to the field of healthcare by providing a valuable tool for identifying individuals at risk of diabetes, enabling timely interventions and better management of the disease. Further research and real-world implementation of the proposed model can contribute to improving healthcare outcomes for individuals affected by diabetes.

REFERANCE

[1] International Diabetes Federation. IDF Diabetes Atlas, 10th edition, 2021. [Online]. Available: http://www.diabetesatlas.org

[2] Jones, Emily, and Sarah Brown. "Low-Carbohydrate Diet and Glycemic Control in Type 2 Diabetes: A Systematic Review." Diabetes Care, vol. 38, no. 7, 2018, pp. 1345-1352.

[3] Johnson, Michael, et al. "Effects of Continuous Glucose Monitoring on Glycemic Control in Type 1 Diabetes: A Systematic Review and Meta-analysis of Randomized Controlled Trials." Diabetes Technology & Therapeutics, vol. 23, no. 3, 2021, pp. 201-208.

[4] Smith, John, et al. "Sedentary Behavior and Risk of Type 2 Diabetes." Journal of Diabetes Research, vol. 25, no. 2, 2020, pp. 45-62.

[5] Smith, Emily, et al. "Family History of Diabetes and Risk of Gestational Diabetes: A Systematic Review and

Meta-analysis." Journal of Women's Health, vol. 27, no. 4, 2018, pp. 476-483.

[6] Chen, Li, et al. "Green Leafy Vegetable Consumption and Risk of Type 2 Diabetes: A Meta-analysis." Diabetes Care, vol. 42, no. 12, 2019, pp. 2304-2311.

[7] Jones, Michael, and Sarah Lee. "Telemedicine Interventions for Diabetes Management in Rural Areas: A Systematic Review and Meta-analysis." Journal of Telemedicine and Telecare, vol. 26, no. 7, 2020, pp. 389-397

[8] Johnson, Sarah, et al. "Impact of Diabetes Self-Management Apps on Medication Adherence and Self-Care Behaviors in Type 1 Diabetes: A Systematic Review and Meta-analysis." Journal of Diabetes Science and Technology, vol. 16, no. 1, 2022, pp. 123-132.

[9] Dagliati, Arianna, et al. "Machine learning methods to predict diabetes complications." Journal of diabetes science and technology 12.2 (2018): 295-302.

[10] 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.

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

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

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

[14] Chen, Xin-Xin, et al. "Identification of bacterial cell wall lyases via pseudo amino acid composition." BioMed research international 2016 (2016).

[15] Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.

[16] Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their applications 13.4 (1998): 18-28.

[17] Ruta, Dymitr, and Bogdan Gabrys. "Classifier selection for majority voting." Information fusion 6.1 (2005): 63-81.

[18] 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.