Ensemble Learning for Diabetes Prediction: A Comprehensive Approach

DR. KAVITA D. HANABARATTI[1], ASHWINI YADRAMI[2], MADHUSHREE METI[3], SAHANA SUNKAD[4], SANJANA KALYANKAR[5]

[1, 2, 3, 4, 5] *Department of Computer Science and Engineering, Karnataka Law Society's Gogte Institute of Technology, Visvesvaraya Technological University, Belagavi, India*

***Abstract*— Diabetes mellitus continues to be a major global health concern, underscoring the vital necessity of precise and timely prediction models. The goal of this project is to create a powerful predictive tool that uses ensemble machine learning techniques to categorize people who are at risk of diabetes based on important health indicators like age, family history, BMI, and blood glucose levels. Developing a complete solution that includes phases for data preprocessing, feature selection, model training, evaluation, and deployment is the aim. This work aims to achieve high accuracy, interpretability, and generalizability in diabetes prediction by utilizing an ensemble approach with a combination of Random Forest, Decision Tree, and Support Vector Classifier models. The ultimate aim is to provide healthcare practitioners with a trustworthy device for early discovery and negotiative techniques, therefore improving patient outcomes and reducing healthcare costs associated with diabetes management.**

***Index Terms*- Ensemble Machine Learning, Diagnostic analytics, Max voting, SVM, Random Forest classifier, Decision tree, Classification report**

I. INTRODUCTION

Diabetes mellitus remains a prevalent health concern worldwide, with its incidence steadily rising and its associated complications imposing a significant load on healthcare systems. Timely discovery and negotiation are critical for effective management and prevention of diabetes-related complications. In recent years, machine learning techniques have garnered attention for their potential to improve the precision and efficiency of diabetes prediction models.

The objective of the provided code is to implement a machine learning-based approach for diabetes prediction, leveraging ensemble techniques to enhance predictive accuracy and robustness. The implementation encompasses various stages, which includes data preprocessing, model selection, training, evaluation, and deployment, with a focus on achieving high performance and usability in clinical settings.

The code utilizes the Pima Indians Diabetes dataset, a widely used benchmark dataset containing health

parameters such as blood glucose levels, BMI, age, and family history. It employs ensemble learning techniques, combining the predictive power of multiple classifiers, containing Random Forest, Decision Tree, and Support Vector Classifier (SVC), to create a comprehensive predictive model.

Through rigorous evaluation using classification reports, the code provides insights into the performance of individual classifiers and the ensemble model, enabling healthcare professionals to assess the reliability and effectiveness of the predictive tool. Furthermore, the implementation facilitates the deployment of the trained model for real-world use, empowering healthcare providers with a valuable device for early discovery and negotiation strategies in diabetes management.

By providing a practical implementation of diabetes prediction using machine learning ensemble techniques, the code contributes to advancing the field of healthcare analytics and underscores the capability of machine learning in improving patient outcomes and reducing healthcare costs associated with diabetes management.

II. LITERATURE REVIEW

This literature review focuses in providing a comprehensive overview of recent advancements in

diabetes prediction using machine learning techniques.

In [1], the authors main objective was to lessen the risk of acquiring diabetes by making forecasts for them and telling them to maintain proper diets and lifestyle.

The prediction should be more accurate and reliable. Therefore the authors have used an ensemble learning algorithm including KNN, label encoder and train test split. This may help the medical professionals to predict more accurately that will help in managing the diabetes.

In[2], the authors have focused mostly on type 2 diabetes. They have applied the eight-parameter logistic regression classification approach on the PIMA Indian Diabetic Dataset. In order to create a new model using the predictions from the base model, the stacking technique is applied.

Machine learning algorithms like Naive Bayes, Support Vector Machine Classifier and Decision tree are used to build the model.

In[3], using the PIMA diabetes dataset, the authors tested five different boosting techniques. To improve the evaluation of the quality of the dataset, exploratory data analysis was applied. XGBoost, CatBoost, LightGBM, AdaBoost, and Gradient Boosting were the several boosting algorithms that were used.

In[4], the authors implement prediction of diabetes mellitus through machine learning. They have used datasets form various different resources like countries from Korea, Bangladesh and the PIMA Indian diabetes dataset.To forecast missing samples, a preprocessing method based on polynomial regression was employed.

In some of the publications, the applicable machine learning models have been integrated into a website or smartphone application.

In[5] the author has done comparative analysis based on accuracy. She has used a correlation matrix to check the relation between the outcome and any of the metrics.The author has used K-nearest neighbor, Logistic regression, Decision tree, Random forest and Support vector machine and also the feature importance in Decision tree and random forest. Finally she compared all of their accuracies with decision tree being the most accurate with accuracy of 98% for training data and 99% for testing data followed by Random forest with 94% and 97% in training and testing dataset respectively.

In[6], the author has mainly focused on review of machine learning in diabetes detection. Here the author talks about machine learning as well as deep learning in diabetes prediction. the author has also talked about algorithms such as ordering points to identify cluster structure(OPTICS).

Therefore, the naïve Bayes (NB) data mining technique is applied, and BIRCH and OPTICS are utilized to identify the optimal algorithm for improved accuracy and to cluster comparable types of data. One of the health analysis systems with the quickest growth is Apache Spark. They have also discussed the operation of irido diagnosis.Also deep learning is quite useful in diabetes detection byusing different types of networks like artificial neural networks, deep neural networks, convolutional neural network and recurrent neural network. Here DNN achieved the highest accuracy of 98%.

Deep learning performs better on most datasets, thus in order to achieve optimal accuracy and performance, it should be paired with other techniques. Hybrid methods have the function of enhancing the models' performance.

In [7], The author has focused mainly on cases in in developing countries and the accuracy with which the diabetes is being predicted. They have proposed a model which includes 5 modules namely:1. Dataset collection- which consists of 800 records and 10 attributes.2. Data Pre- processing- it is done to handle inconsistent data like missing values.3. Clustering: In this the author has used K- means clustering on the dataset to classify each patient into either a diabetic or non-diabetic.4. Build model and5.Evaluation The author has used algorithm 1 as Diabetes prediction using various machine learning algorithm and also the algorithm 2 is Diabetes Prediction using pipeline.

In the results it was found that Logistic regression has the highest accuracy of 96% followed by LDA with accuracy of 94% and so on. Finally by using pipeline AdaBoost classifier gave the best model accuracy of 98.8%.

In[8] A dynamic predictive modeling approach for real- time chronic illness prediction has been proposed by the authors. There are four basic phases to the approach. Phase 1 focuses on gathering data from various sources and creating a consolidated patient unique identifier indexed database on a cloud platform. The dataset is constantly expanding. By taking disease-specific parameters out of the centralized tables, phase 2 disease-relevant sub-tables are created. Amputations of peak values and missing values are preprocessed in the sub-tables. Phase 3 involves updating the chosen model's parameters on the server that has the learned model. The server also stores the trained machine learning models, which are updated on a regular basis after a time interval of 'T'. Phase 4 involves adding a new, unclassified test record with a unique PUID to the centralized database. This record serves as test data and is processed by the trained model. The author has utilized Random Forest ML methods, K-nearest neighbor algorithms, Naive Bayes, Decision trees, Neural Networks, Logistic Regression, and Support Vector Machines. For each algorithm, the classification result and confusion matrix are provided. Additionally, all seven learned models' ROC curves are plotted at optimal hyperparameter settings. The SVM and neural network models have the highest prediction accuracies, indicating their extreme dependability.

In[9], Using a private dataset of diabetes mellitus is the author's primary contribution, as stated in the paper itself. Three hundred volunteer data samples that were collected from the hospitals make up the dataset. This work also makes use of the Synthetic Minority Oversampling Technique (SMOTE) to effectively manage the imbalanced classes. In addition, the research assesses ten distinct machine learning categorization methods in order to identify the system that generates the best precise diabetes forecast. All things considered, this work advances the field of ML-based diabetic prediction with new methods and insights.

In [10], the author states that crucial variables including frequent urination, weight loss, frequent feeding, vaginal thrush, hazy vision, sluggish healing, and paresis must be present in the dataset in order to create the optimal diabetes prediction model. Nineteen machine learning classification methods were used in the development of the diabetes prediction model. K-fold cross-validation was used to repeatedly test the top machine learning algorithms, GBoost, LightGBM, and random forest, to ensure that the model was stable. The effectiveness of the model was tested using a variety of metrics, such as accuracy, precision, recall, and F1-score.

The ROC curves and the confusion matrix were both used in the model's efficacy assessment. The study's conclusions demonstrate the model's precision and efficacy in estimating the risk of diabetes, offering insightful information to both researchers and medical professionals.

In[11], The project's primary goal was to successfully design and execute diabetes prediction using machine learning techniques and performance analysis of those techniques. The suggested strategy makes use of a variety of ensemble learning and classification techniques, including SVM, Knn, Random Forest, Decision Trees, and Logistic Regression. And the KNN algorithm has obtained 78.5% classification accuracy. The outcomes of the experiment can help medical professionals foresee the future and make decisions early on to treat diabetes and save lives.

In[12], In order to forecast diabetes from the patient data set, they have used various machine learning method classifiers in this research. They were able to use various machine learning methods to produce various classifier models. They were able to determine the model with the highest accuracy thanks to this investigation. The Random Forest model was the most precised of all of the models.

In[13], By using the proposed ML-based ensemble model, which emphasizes the importance in providing reliable and accurate predictions, this research was able to accomplish its goal of developing an accurate and reliable diabetes prediction. The preprocessing method that was described increased the quality of the data; choosing features and Elfing in moving values

were crucial factors. The application of these preprocessing techniques in Nevine required an analysis of the ablative processes to select the most appropriate ways. Furthermore, in comparison to earlier research, this study yields a more precise characterization that comprises only four to five variables: the respondent's body mass index (BMI), their age, their average age systolic pressure, their average diastolic pressure.

In[14], Significant progress has been made in a number of domains, including Device Learning Database Systems and Simulated Intelligence. This PIMA dataset is mare. The main goal is to improve the predicted accuracy of all algorithms, however SVM and linear regression perform better than others. Using a variety of cutting-edge methods, the wood will source the algorithms' efficiency.

In[15], The author has studied the algorithm for support vector machines Thus, in supervised machine learning, the machine learning model learns from the analysis of the corresponding label after pre-processing the data with training. This data must be standardized so that it can be used for training and tasting. After that, we try to gauge how well our model is working based on the analysis we have conducted and have established a standard score for it. To do this, we must pre-process the data using the same range machine learning model that Pima Indian diabetes dataset uses.

In[16], Based on accuracy, a few current Ml. CDAO canon models for the prediction of diabetes patients have been explored. The classification problem has been recognized as an expression of accuracy. Healthcare Engineering Journal The FIDD data set was used to further refine the technique. On the dataset and vert, it was trained and verified. The results demonstrate how the outperformed alternative MI algorithms. It has been discovered that the KOC value of is 86% and that glucose and BMI have a high correlation with diabetes wings association. The drawback of this work is that unstructured data will be taken into consideration for the analysis of the structured data set that we have chosen.

In[17], The potential of machine learning approaches to revolutionize early diabetes diagnosis and care is thoroughly reviewed in this research. The findings show that different machine learning algorithms can reliably forecast the onset and risk of diabetes when they are used with a variety of health data sources. On several datasets, models with over 90% precision, such as random forests, XGBoost, and neural networks, demonstrated excellent performance. To fully reap the rewards of these advancements in clinical care, a number of important obstacles must be overcome. Thorough real-world validation guarantees effectiveness across a range of demographics and prevents inadvertent bias. Adoption will depend on a seamless interface with electronic data and clinical operations. Building more interpretability and transparency into the model will increase clinician comprehension and trust. Standardized, high-quality data is the cornerstone of accurate forecasting.

In[18], The main objective of this study was to compare the efficiency logistic regression with other linear classifiers, such as random forest (RF), SVM, and KNN. The comparison's findings showed that Logistic Regression fared better than every other classifier. At 0.83%, it was discovered that the accuracy of logistic regression was the highest. High accuracy levels were obtained by using ensemble learning and classification techniques in the suggested strategy. By enabling early predictions and informed decisions, these experimental results can help medical professionals. These classifiers can also be used to cure diabetes and may even save human lives.

In[19], Several machine learning techniques are employed in this study to classify the dataset; Random Forest yields thegreatest accuracy of 90% among these algorithms. In order to minimize False Negative values, we have observed comparisons of confusion matrices and machine learning method accuracies. By determining whether or not the non-diabetic individual is likely to develop diabetes in theupcoming years, we can further our research.

III. PROBLEM STATEMENT

The problem of diabetes prediction using machine learning involves developing a model that can accurately classify whether an individual is likely to have diabetes

based on their health parameters such as blood glucose levels, BMI, age, family history, etc. The goal is to develop a predictive tool that will help medical practitioners identify problems early and implement intervention plans that will ultimately improve patient outcomes and save medical expenses. The goal of this problem is to achieve high accuracy, interpretability, and generalizability of the predictive model through the steps of data preparation, feature selection, model training, evaluation, and deployment.

IV. DETAILED METHODOLOGY

4.1 Dataset description

The Pima Indians Diabetes dataset is a conventional dataset used under machine learning ground for predicting diabetes occurrence in Pima Indian populationbased on certain diagnostic measures. This dataset consistsof 768 instances, each with 8 attributes such as the number of pregnancies, plasma glucose concentration, blood pressure, skin fold thickness, serum insulin, body mass index (BMI), diabetes pedigree function, and age, along with a binary target variable marking the presence or absence of diabetes. The dataset is valuable for exploring various machine learning algorithms and techniques for classification tasks, as it presents challenges such as class imbalance and missing values. Researchers and practitioners often utilize this dataset to develop and evaluate predictive models to assist in diabetes diagnosis and management, contributing to advancements in healthcare analytics and personalized medicine. To enhance the robustness of the models and improve prediction accuracy, we have used bootstrapping method. Bootstrapping involves resampling data from the original dataset with replacement to create multiple bootstrap samples which has the same size as the original dataset but may contain duplicate instances due to sampling with replacement.

4.2 Proposed Model

Collection and Preprocessing:

Gather a dataset containing features relevant to diabetes prediction such as glucose level, blood pressure, BMI, age, etc.

Preprocess the data by serving missing values, scaling numerical features, and encoding categorical variables if any.

Model Training:

Train individual models:

- Random Forest: Ensemble learning method that constructs multiple decision trees and merges them to get a more precised and stable prediction.
- Support Vector Machine (SVM): A supervised learning algorithm that can categorise data by identifying the hyperplane that differentiate among classes.
- Decision Tree: A tree-like model where an internal node represents a feature, the branch represents a decision rule, and each leaf node represents the outcome.

Ensemble Model Creation:

Utilizing a Max Voting ensemble approach, aggregate the predictions from each individual model. Using this procedure, all base model predictions are combined, and the class with the most votes is chosen as the final forecast.

Evaluation:

Evaluate the performance of each individual model and the ensemble model using relevent evaluation metrics such as accuracy, precision, recall, F1-score, and support

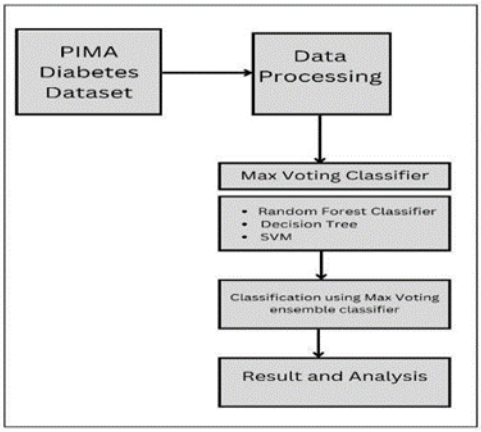

Figure 1: The flow diagram of proposed ensemble model using max voting classifier.

V. RESULTS AND CONCLUSION

In this study, we used machine learning techniques to predict the likelihood of diabetes occurrence in individuals based on several health-related features. We explored the effectiveness of ensemble learning,

combining multiple machine learning algorithms, namely Random Forest, Support Vector Machine (SVM), and Decision Tree, to enhance prediction accuracy.

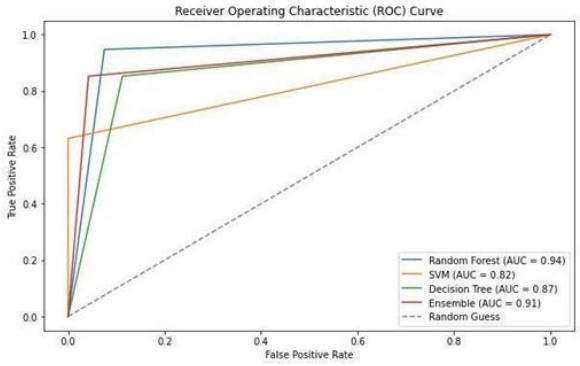

Figure 2: Receiver Operating Characteristics (ROC) Curve for different classifiers

ALGORITHMS	ACCURACY	SUPPORT	PRECISION	F1_SCORE	RECALL
SUPPORT VECTOR	90.0	95.0	100.0	77.0	63.0
DECISION TREE	88.0	95.0	77.0	81.0	85.0
RANDOM FOREST	94.0	95.0	85.0	90.0	95.0
MAX-VOTING	92.0	95.0	90.0	88.0	85.0

Figure 3. Comparison of various Machine Learning models

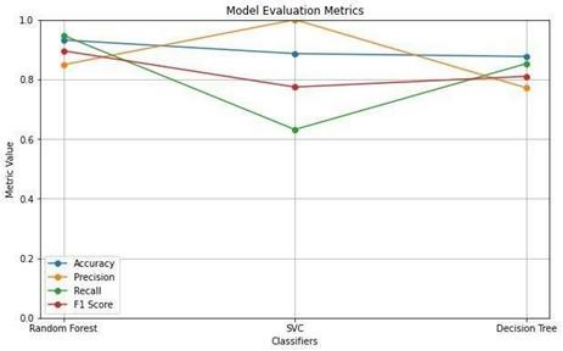

Figure 4. Graph showing Comparison of various Machine Learning models

The study's findings demonstrate the potential of Max Voting ensemble approaches as a useful instrument for clinical practice's diabetes prediction. Healthcare professionals may get more accurate risk assessments by combining the expertise of several machine learning models. This allows for early intervention and individualized treatment plans for people who are at high risk of getting diabetes.

VI. FUTURE WORK

In Future Work we can experiment with different ensemble models and find which is more efficient in terms of not only accuracy but which utilizes less time to process and also uses the least memory without compromising on the accuracy or other metrics. If the dataset suffers from class imbalance, consider techniques such as oversampling, undersampling, or using different evaluation metrics to address this issue. If the dataset is too large we can explore on optimizing the code and also we can develop user friendly interface for the common people to use it conveniently

VII. ACKNOWLEDGMENTS

The department of Computer Science and Engineering, KLSGIT, Belagavi is providing support for this paper. We would like to extend our sincere and heartfelt thanks towards the Department as well as to the university.

REFERENCES

[1] G. Parimala, Ramalingam Kayalvizhi, S. Nithiya, “Diabetes Prediction using Machine Learning”, in ICCCI, 2023.

[2] C.S. Manikandababu, S. IndhuLekha, J. Jeniefer, T. Annie Theodora., “Prediction of Diabetes using Machine Learning”, in ICECAA, 2022.

[3] Shahid Mohammad Ganie , Pijush Kanti Dutta Pramanik, Majid Bashir Malik , Saurav Mallik , Hong Qin, “An ensemble learning approach for diabetes prediction using boosting techniques.” by PMC, 26 Oct 2023.

[4] Isfafuzzaman Tasin, Tansin Ullah Nabil, Sanjida Islam, Riasat Khan, “Diabetes prediction using machine learning and explainable AI techniques”, in Healthcare Technology Letters, December 2022.

[5] KM Jyoti Rani, "Diabetes Prediction Using Machine Learning”, in International Journal of Scientific Research in Computer Science Engineering and Information Technology, July 2020

[6] Toshita Sharma, Manan Shah, “A comprehensive review of machine learning techniques on diabetes detection”, by PMC,

[7] Aishwarya Mujumdar, Dr. Vaidehi V, "Diabetes Prediction using Machine Learning Algorithms", International Conference on recent trends in Advance Computing 2019, ICRTAC

[8] Nidhi Arora, Shilpa Srivastava, Ritu Agarwal, Vandana Mehndiratta, Aprna Tripathi, "Diabetes mellitus prediction using machine learning within the scope of a generic framework", in Indonesian Journal of Electrical Engineering and Computer Science, December 2023.

[9] Hosam El-Sofany ,SamirA. El-Seoud ,OmarH.Karam,YasserM.Abd El-Latif , Islam A. T. F. Taj-Eddin, "A Proposed Technique Using Machine Learning for thePrediction of Diabetes Disease through a Mobile App", in International Journal of Intelligent Systems Volume 2024.

[10] Karthick Kanagarathinam, Manikandan Radhakrishnan, T Sathish Kumar, "Machine learning algorithms-based decision support model for diabetes", in Review of Computer Engineering Research , January 2024.

[11] Miss. Vaishnavi Khalate', Prof. Borate Sukeshkumar Mar, "Diabetes Prediction Using Machine Learning Algorithm", in IJRASET, 2024.

[12] Melbin Varghese and Mrs. Nimmy Francis, "Diabetes Prediction Using Machine Learning Algorithms", in NCECA, 2021

[13] Aishwariya Dutta, Md. Kamrul Hasan, Mohiuddin Ahmad, Md. Abdul Awal Akhtarul Islam, Mehedi Masad and Hossam Mesheet, "Early Prediction of Diabetes Using an Ensemble of Machine Learning Models", in International Journal of Environmental Research and Public Health, 28 September 2022.

[14] Dr.0.Obulesu, Dr.K.Suresh, B. Venkata Ramudu, "Diabetes Prediction using machine learning techniques", in Helix – The Scientific Explorer, 30 April 2020

[15] Sania Faraz1, Pawan Singh2, "Diabetes Prediction using Machine Learning", in Journal of Applied Science and Education, 2022.

[16] Raja Krishnamoorthi, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C. Kalpana, and Basant Tiwari, "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques", in Journal of Healthcare Engineering, 24 May 2023.

[17] Dinesh Kalla, Nathan Smith, Fnu Samaah and Kiran Polimetla, "Enhancing Early Diagnosis: Machine Learning Applications in Diabetes Prediction", in Journal of Artificial Intelligence & Cloud Computing, 2022.

[18] Ms. P. V. Deshmukh, Ashwini Ghate, Prajakta Mathe, Aditi Dhote, Pratiksha Patte, Vrushali Mange, "Diabetes Prediction using Machine Learning", in International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), April 2023.

[19] Rishab Bothra, "Diabetes Prediction using Machine Learning Algorithms", in International Journal of Engineering Applied Sciences and Technology, 2021.