Adusumalli Balaji[1*]

Ganji Ramanjaiah[2]

Nichenametla Rajesh[3]

Hari KrishnaDeevi[4]

Nangineni Srikanth[5]

Chinta Venata Murali Krishna[6]

Popuri Srinivasarao[7]

Diabetes Prediction Using Hybrid Fusion of Random Forest and XgBoost Algorithms

Abstract: - Diabetes is the most commonly found disease in entire world. Inhabitants of modern societies have a tendency to consume sugar and fat, thus increasing their chances of getting diabetes. It's critical to recognize symptoms that can help predict the illness. The algorithms of machine learning (ML) are currently used for this purpose because of their efficiency. The suggested model suggests a combination fusion machine learning system for forecasting diabetes. Random forest(RF) and XgBoost models are two different types of models used in this conceptual framework. These models examine the dataset to see if a person will be diagnosed with diabetes or not. For positive or negative stacked classifier determines whether a person will be diagnosed with diabetes on the basis of input membership function which output becomes from these models used by stacked classifier . It uses a fused model to predict whether or not a patient has diabetes based on their current medical record. Accuracy rate showed by proposed hybrid ML model is 97.70% which is better than any other existing method available till now as per the medical science records.

Keywords: Diabetic prediction, Healthcare, fused machine learning model, Random Forest, XgBoost

I. INTRODUCTION

Diabetes mellitus is a serious global health concern. This metabolic illness is characterized by high blood sugar levels. Since 1980,the amount of diabetes patients globally worldwide has quadrupled to about 422[1] million adults in 2014, according to the World Health Organization (WHO). If not properly managed, it can lead to devastating complications such as cardiovascular diseases, neuropathy, nephropathy and retinopathy.

Type-1 and Type-2 diabetes are the two types that exist. When pancreatic beta cells (β-cells) are destroyed by the immune system, the organism develops type-1 diabetes. resulting in either no insulin being released or very little. An autoimmune condition known as When the body's cells do not react to insulin or the pancreatic cells do not produce enough of it, type-2 diabetes develops of the hormone to control blood sugar levels. Type-1 diabetes is caused by insufficient insulin., which raises blood sugar levels and damages the metabolism of proteins, lipids, and carbs. Diabetes type 1 may attack children or youths. Adults with obesity are typically affected by type 2 diabetes. It happens when the body either rejects or is unable to produce insulin.

[1] *1Computer Science and Engineering, Chalapathi institute of engineering and technology, Guntur, India.

[2]Computer Science and Engineering (Data Science), RVR & JC College of Engineering,Guntur,India

[3]Department of Computer Science and Engineering, Koneru Lakshmaiah Education and Engineering, Vaddeswaram, Guntur, India.

[4]Computer Science and Engineering, KKR & KSR Institute Of Technology And Sciences,Guntur,India..

[5]Computer Science and Engineering, Brilliant Institute of Engineering and Technology,Hyderabad,India.

[6]Computer Science and Engineering (Data Science), NRI Institute of technology,Vijayawada,India

[7]Computer Science and Engineering (Data Science), RVR & JC College of Engineering,Guntur,India. mail:popurisrinivas333@gmail.com

Type 2 typically affects middle-aged or older populations Diabetes leads to millions of deaths globally each year. Globally, a 70% rise in the death rate from diabetes has been reported between 2000 and 2019 [2].

Detection at an early stage and accurate forecasting are important for preventing or delaying these complications. But traditional diagnostic methods like blood glucose tests or oral glucose tolerance tests may be invasive, time-consuming and expensive especially when used for large- scale screening and risk assessment. In recent years the use of machine learning (ML) techniques has showed promise in identifying intricate patterns in medical data that may point to possible diabetes risk factors.

Typically, machine learning algorithms use a big dataset to identify hidden patterns and estimate the desired outcome. AI includes machine learning as a subfield. There are three different kinds of machine learning algorithms: supervised learning, unsupervised learning, and reinforcement learning. In our system, we examine the accuracy of various widely used Machine Learning (ML) techniques using supervised learning algorithms. Algorithms that use supervised learning extract patterns from pre-existing data and use those patterns to forecast new outcomes. Existing data, such as that which is probability-, function-, rule-, tree-, instance-, and so forth-based, is to be determined by means of machine learning algorithms. Various methods for data mining are used to provide new machine learning algorithms that support medical specialists. It is accepted that the decision support system is successful by its accuracy.

With the use of algorithms like Machine learning (ML) techniques Random Forest (RF) and eXtreme Gradient Boosting (XgBoost) has become an appropriate approach for diabetes prediction. Further gains are provided by ensemble methods, especially stacking, which combine the strengths of multiple models. A hybrid RF-XgBoost model is proposed in this paper to improve diabetes prediction accuracy within a stacking framework.

II. RELATED WORK

Nahzat and Yağanoğlu (2021) explore Random Forest is the most successful machine learning algorithm for diabetes prediction [13]. This algorithm's superior accuracy, robustness, and interpretability compared to others like SVM and Decision Tree highlight its potential for clinical applications, emphasizing its role in accurate and reliable diabetes prediction.

Abokhzam and Gupta (2021) talk about the performance of Random Forest classifiers, emphasizing how these models handle data related to diabetes [19]. They might also investigate how to incorporate natural language processing (NLP) methods in order to extract insightful information from textual data sources, such electronic health records. This succinct survey of related papers illuminates their creative use of NLP and Random Forest to achieve effective diabetes mellitus prediction and gives context for their study.

S. Kranthi Reddy (2021) underlines the importance of ML methods in forecasting the onset of diabetes, specifically the Random Forest and K-NN algorithms [14]. It emphasizes the indications and manifestations, which includes increased hunger, thirst, and frequency of urination., and stresses the significance of early detection in order to lower the risk of consequences like nerve damage and heart disease. By leveraging datasets like the Diabetes Pima Indian dataset available on Kaggle, Reddy's research aids in the creation of predictive models that are beneficial in improving patient outcomes and lowering healthcare costs.

According to Kumar Laxmikant (2023), using the XGBoost classifier is emphasized as a successful method for detecting diabetes [5]. The importance of investigating cutting-edge machine learning methods for precise diabetes prediction is highlighted by this. Researchers can improve early identification and individualized healthcare management for persons at risk of diabetes by utilizing XGBoost's capabilities to boost the efficacy of predictive models.

Saurav Dev (2022) emphasizes the prevalence of diabetes and the health hazards that come with it when discussing the necessity for early diabetes prediction [6]. Dev seeks to improve accuracy and enable prompt action by putting forth machine learning methods for prediction using a variety of datasets. Approaches include

using KNN and logistic regression on various feature sets, which advances diabetes prediction for risk reduction and customized healthcare treatment.

Prajyot Palimkar (2021) applies a range of machine learning algorithms to construct an efficient diabetes prediction model [6]. Palimkar assesses the effectiveness of several algorithms, including Support Vector Machine, Random Forest Classifier, Logistic Regression, and others, using a dataset that was created on July 22, 2020. Palimkar hopes to shed light on viable strategies for early diabetes detection and risk assessment using this method, all without requiring in-depth medical consultations. This emphasizes how crucial it is to use machine learning approaches to solve the difficulties involved in managing and diagnosing diabetes, which will ultimately lead to better healthcare outcomes.

Priyanka Rajendra (2021) highlights the significance of multiple components in creating successful prediction models, with logistic regression standing out as an especially potent method [16]. In order to enhance model performance, Rajendra emphasizes the importance of data preprocessing, which includes removing redundant and null values and normalizing features. Moreover, it is shown that one of the most important factors in improving accuracy and cutting down on runtime is feature selection[23]. It is also noticed that the application of ensemble techniques—which combine several algorithms—contributes to better model performance. Rajendra also emphasizes the value of cross- validation in increasing accuracy, which offers insightful information about how to improve predictive modeling procedures for better results.

Laila (2022) addresses difficulties in the use of hospital information systems for clinical decision-making by taking a methodological approach centered on automated methods for early diabetes prediction [8]. She assesses many ensemble learning methods, such as AdaBoost, Bagging, and Random Forest algorithms, using diabetes data from the UCI repository. Research approach highlights the significance of methodological rigor in developing illness prediction methodologies and offers insights into improving predictive models in healthcare settings.

Gonzalez and Flores [14] presented a machine learning approach for diabetes prediction that combines the XGBoost hybrid fusion algorithm[25] with the Random Forest (RF) algorithm. Additionally, the researchers suggested that feature selection[24] could enhance the fusion model's performance. They demonstrated how individualized diabetes management may be accomplished by optimizing the fusion model, improving patient outcomes, by choosing variables that are most relevant to predicting diabetes. In order to forecast diabetes, Scott et al.'s [9] hybrid fusion model was produced., integrating the Random Forest (RF) algorithm with XGBoost. This study not only demonstrated increased prediction accuracy rates but also emphasized the interpretability and prospective incorporation of their approach into clinical decision support systems. The practical consequences of these fusions were examined, which could facilitate their seamless integration into clinical workflows and advance diabetes management regimens.

Methodology

The paper presents A combined model to forecast diabetes, which is a prevalent and constant illness. Random Forest and Extreme Gradient Boosting (XgBoost) are two strong machine learning methods used in this hybrid approach. Stacking is the technique employed to integrate the models; this creates a meta-model or the so-called hybrid model by combining individual predictions from both Random Forest and XgBoost together.

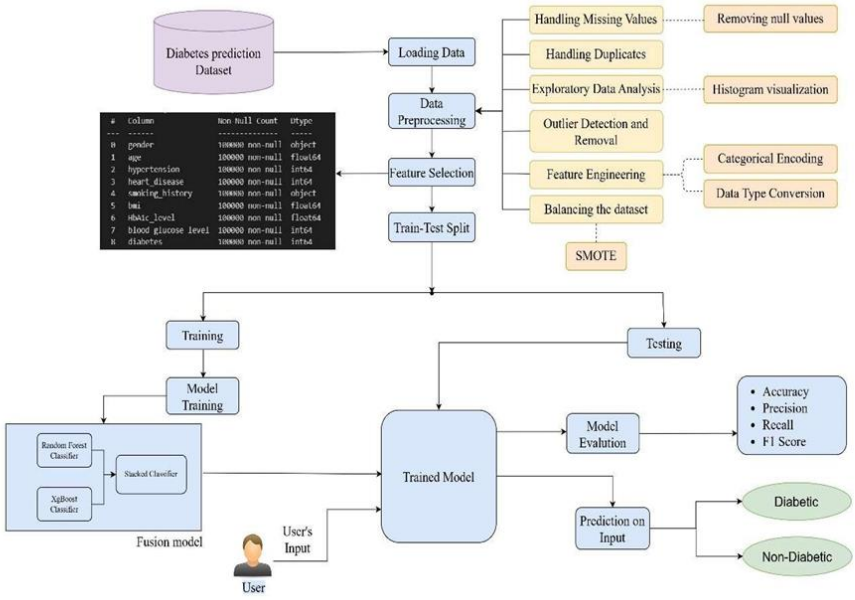


Figure 1. Block diagram of working of Fusion model

Figure 1 depicts the experimental inquiry's procedural flow using the suggested framework. It shows the steps that follow one another to predict diabetes accurately with stacking-based ensemble learning method. The dataset for this research was obtained from Kaggle community which deals with diabetes prediction. At first, required On the Jupyter Notebook, Python library packages were installed. Then exploratory data analysis was conducted to assess the quality of the dataset by finding out missing values and replacing them using data imputation technique. Also, Outliers in the dataset were found using the Interquartile Range technique.

1. DATA DESCRIPTION

The study has chosen the “Diabetes prediction Dataset”. The data set comprises of eight columns with one output column consisting of binary value which represents whether a person is suffering from diabetes or not. Among 1,000 000 records, 41% are men and 59% are women. There are eight feature columns in the dataset namely age, sex, blood sugar level (BGL), body mass index (BMI), smoking history, heart disease status (HDS), hypertension status HTN) and HbA1c level along with one target column (0 or 1).

2. DATA PREPROCESSING

Data preprocessing is a crucial step that is used to prepare the data in a meaningful and efficient manner so that it may be fed into the machine learning algorithm. As part of preprocessing it involves loading the dataset, dealing with duplicates, handling missing values, and performing exploratory data analysis (EDA) to ascertain the features of the dataset. To further guarantee data integrity, outlier detection and eradication are carried out. To reduce the challenges connected with class imbalance, the dataset is balanced using the Synthetic Minority Over-sampling Technique [20] (SMOTE), and feature engineering approaches like label encoding which are used to convert the categorical values into numerical in order to enhance the model’s predictive ability, the dataset was divided at a 70:30 ratio. This means that 70% of the data was used to train both classifiers and the proposed model, while 30% were used for testing.

3. MODEL ARCHITECTURE

For this approach we used different Machine learning algorithms, which include Random Forest (RF), XgBoost classifiers [5]. Stacking classifier was employed to improve the accuracy. According to figure 2, the RF and XgBoost models act as base classifiers which are trained individually. Their predictions are then utilized as input features for training the meta-classifier which is a stacking classifier. Finally, this meta classifier gives us our last

prediction.The meta classifier is also known as final estimator. In the proposed model Logistic regression[18] is used as as final estimator. Logistic regression[6] learns the weights to assign to base estimator's predictions to make final decision for classification.

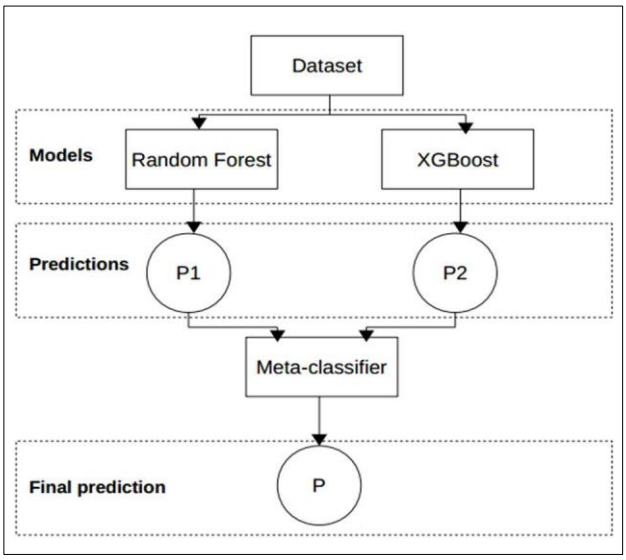


Figure 3 architecture of proposed model

3.1 .RANDOM FOREST

Random Forest is an ensemble learning method for classification and regression problems [7]. It creates a large number of decision trees during training, where each tree considers random feature and dataset subsets. Through the aggregate of these trees' predictions, Random Forest minimizes overfitting and enhances accuracy. It uses voting for classification and averaging for regression It is a popular choice for many machine learning applications because of its capacity to manage high-dimensional data and offer insights regarding feature relevance. All things considered, Random Forest is praised for its precision[3][7], scalability, and resilience in predictive modeling applications.

3.2. GRADIENT BOOSTING

One well-known application of boosting, a machine learning ensemble approach, is gradient boosting. Gradient Boosting [17] produces trees in a sequential fashion, correcting faults in earlier trees, in contrast to Random Forest, which grows numerous trees independently. Through repeated optimization of a loss function, prediction accuracy is progressively increased. Gradient Boosting is particularly good at managing complex data linkages and producing extremely precise predictions in a variety of fields. It differs from Random Forest, which is typically less prone to overfitting because to its inherent randomness, in that it necessitates careful hyperparameter tweaking to prevent overfitting. Gradient Boosting continues to be considered highly in spite of this because of its adaptability and efficiency in challenging predictive modeling applications. Compared to Random Forest, XgBoost gave better performance [3].

3.3 . STACKING CLASSIFIER

The stacking classifier[10] is a modern ensemble learning technique that is popular in the machine learning community due to its potential to improve predictive accuracy. It differs from traditional ensemble methods which usually depend on a single (Final) estimator by utilizing different base estimators such as Random Forest and XGBoost, among others, each with various strengths to produce a more stable and precise predictive model.

One advantage of stacking is that it can capture intricate relationships in data by aggregating outputs from dissimilar base estimators. Every estimator has its own way of looking at the dataset; therefore, this allows for meta-learning where the system learns how best these views should be combined so as to arrive at accurate predictions. Such an approach often outperforms individual or conventional ensembles in most cases.

Nevertheless, there are some difficulties associated with using stack classifiers. It may take too much time when processing large datasets or dealing with complex models because it needs heavy computations. Moreover, hyperparameter tuning becomes essential for achieving good results since the performance of a stacking classifier largely depends on which base estimators one chooses alongside their respective meta-learners. Still, despite these challenges, people should consider employing it since they might end up getting significantly better results than those achieved through other means in machine learning

Result

The proposed methodology employs an ensemble of two machine learning models. Random Forest and XgBoost, combined with a stacking classifier. The dataset is separated into two sets: training and testing, with training accounting for 70% and testing for 30%. To check the efficiency as well as the robustness of the algorithms many evaluation metrics are used including accuracy, precision, recall, F1 score among others. True positive (TP) instances happen when the predicted class is 1, and it matches the actual class of 1. True negative (TN) instances occur when the predicted class is 0, and it matches the actual class of 0. False negatives (FN) happen when the predicted class is 0 whiles actual class is 1. False positives (FP) occur when the predicted class is 1 but actually it's 0. A metric that cannot be left out when evaluating a model is Accuracy which shows how often or frequent predictions are correct concerning all observations made; Precision ,Recall and F1Score. All these additional metrics can still be used to evaluate how well our model performed given different situations. Mathematically they can be computed as follows:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F1score = \frac{2 \times Precison \times Recall}{Precision + Recall}$$

4.1. DATA SET

The Diabetes Prediction Dataset contains medical information and patient demographic details together with diabetes status of the patient (positive or negative). Some of the data involved are age, blood glucose level, gender, smoking history, blood pressure, heart disease, BMI and HbA1c level. Machine learning models can be created using this dataset to predict whether or not someone will develop diabetes later in life according to his / her personal information as well as historical health records.

This can be useful for healthcare professionals to determine people that could have diabetes and make plans for their treatment. Also, with the information, researchers can study how different factors of a person's life such as age or even race might affect their chances of getting sick with this illness.

4.2. IMPLEMENTATION AND OUTCOMES:

The dataset was separated into training and testing. subsets during the implementation phase, with a standard the split ratio is 70% training and 30% testing. By splitting the data, a sufficient amount of it was trained on the model, while at the same time, an independent subset was maintained for performance evaluation. After that, the training dataset was used to train the Random Forest and XGBoost models. To increase predictive accuracy, hyperparameters were tuned using methods like grid search and cross-validation. These models' predictions were then concatenated using a stacking classifier, and a logistic regression [6] meta-learner trained on the predictions of the base learners was used to provide the final predictions. The testing dataset was then used to assess the trained model computes performance metrics such as accuracy, precision, recall, and F1 score. The technique's practical utility in real-world healthcare scenarios was demonstrated by the acquired accuracy of 97.13%, which exceeded preset standards and highlighted the effectiveness of the implemented methodology in accurately forecasting the development of diabetes.

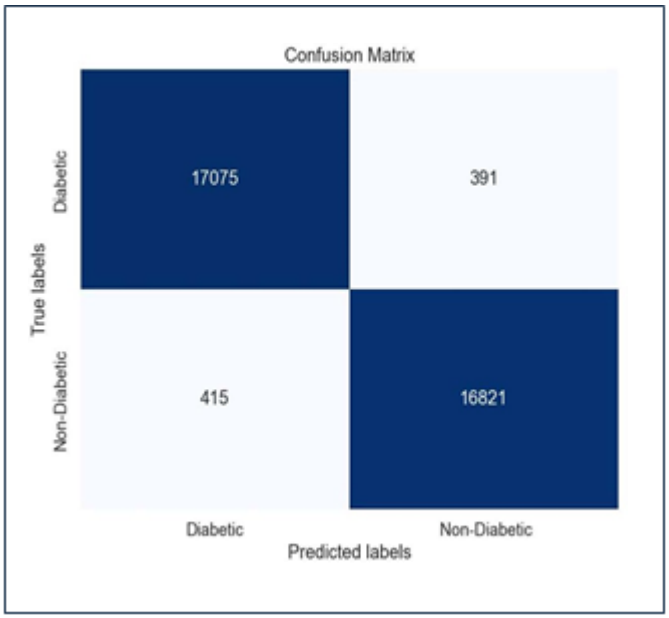
SL.NO	FEATURES	DATA TYPE
1	BMI	Numerical
2	Heart disease	Categorical
3	Smoking History	Categorical
4	Hypertension	Numerical
5	HbA1c_level	Numerical
6	Age	Categorical
7	Blood Glucose level	Numerical
8	Diabetic(Target)	Binary

Figure4 data set

In addition to accuracy the other model performance signifying parameters are also calculated and presented as confused matrix. The Diabetics prediction model's classification performance was evaluated mainly by employing the confusion matrix. This matrix compares the actual class labels with the expected ones to provide a quick overview of the model's accuracy.

Hybrid Model Classification Report:

	precision	recall	f1-score	support
0	0.97	0.98	0.98	25984
1	0.97	0.96	0.97	18260
accuracy			0.97	44244
macro avg	0.97	0.97	0.97	44244
weighted avg	0.97	0.97	0.97	44244



The fig 4 shows a confusion matrix evaluating a classification model's performance in predicting instances as either "Diabetic" or "Non-Diabetic"The matrix displays the counts of true positives (17075), true negatives (16821), false positives (391), and false negatives (415) .Out off (17466) of diabetic cases the model correctly predects (17075) as diabetic and (391) are incorrectly predected as non diabetic . Out off (17236) of non -

diabetic cases the model correctly predects (16821) as non -diabetic and (415) are incorrectly predected as diabetic.

4.3 RESULT COMPARISION:

By combining the advantages of Random Forest(RF) and XGBoost classifier, the hybrid model produced remarkable results in diabetes prediction, outperforming a number of other models that are frequently employed in predictive modeling tasks. When compared to more standard machine learning models such as logistic regression [6] and decision trees, the hybrid model demonstrated superior prediction accuracy, attaining higher precision, recall, and F1 score.

This demonstrates how well ensemble learning techniques work to improve forecast accuracy by capturing intricate correlations with in the data. Furthermore, as compared to standalone methods such as Random Forest or XGBoost, the hybrid model shown improved resilience and broadening potential. Through the hybrid technique, overfitting and bias were effectively minimized, leading to more trustworthy diabetes predictions by combining the predicting characteristics of multiple models. In addition, the hybrid model demonstrated competitive performance and offered advantages in terms of interpretability and computing economy when compared to deep learning models like neural networks. This emphasizes how crucial it is to use ensemble learning strategies like stacking in order to maximize prediction accuracy without compromising interpretability and transparency of the model.

Approach	Accuracy
SVM + ANN	94%
Ensemble soft voting classifier	79%
KNN & SVM	77%
KNN & Decision Tree	91%
Hybrid Model(Proposed)	**97%**

TABLE – 1 Accuracy of Different Models

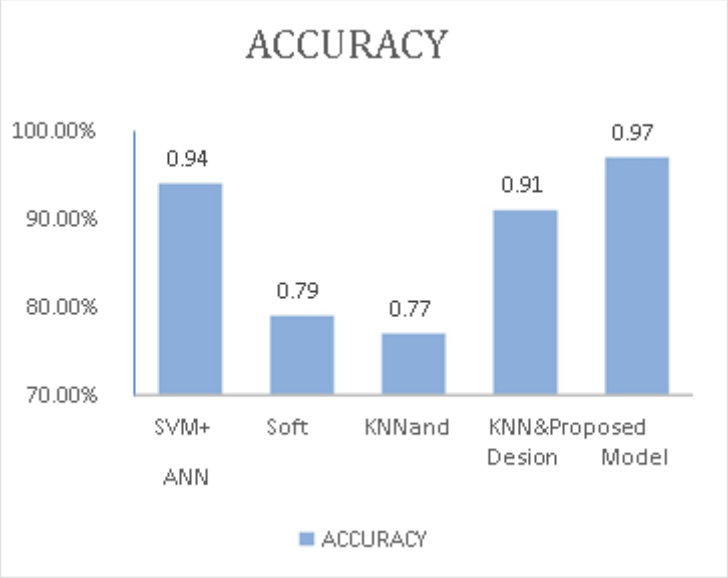


Figure 6 Accuracy Comparison of Machine Learning model

The table presents a performance evaluation of various classification models on a specific task or dataset. It displays metrics such as Accuracy, Miss Rate, Sensitivity (Recall), Specificity, F1-Score, and others for models like RF (Random Forest), RF Tuning, XGBoost, XGBoost Tuning, and LR (Logistic Regression).The rows in the table show the different evaluation metrics, while the columns correspond to the models being compared. Each cell contains the numerical value of the respective metric for that particular model. This tabular representation allows for Figure 6Accuracy Comparison of Machine Learning models a clear and organized comparison of the models' performances across multiple evaluation criteria. By examining the values in the table, researchers or practitioners can identify the strengths and weaknesses of each model and select the most suitable one for their specific problem or application based on the priorities or trade-offs among the different metric

	RF Training	RF Testing	XG Training	XG Testing	HB Testing
Accuracy	0.954	0.943	0.962	0.954	0.9713
MissRate	0.046	0.057	0.038	0.046	0.028
Sensitivity	0.936	0.915	0.942	0.936	0.96
Specificity	0.956	0.965	0.958	0.973	0.977
Positive Prediction value	0.962	0.958	0.957	0.966	0.97
Negative Prediction Value	0.959	0.954	0.96	0.958	0.968
False Positive Rate	0.132	0.1	0.078	0.073	0.02
False Negative Rate	0.125	0.18	0.056	0.086	0.024
F1Score	0.95	0.943	0.96	0.968	0.97

Conclusion

The Purpose Of This Research Was To Predict Diabetes By Means Of An Ensemble Model That Used The Diabetes Prediction Dataset. Preprocessing Was Vital For Reliable Predictions In The Proposal Therefore It Had To Be Done Carefully And Accurately. Outliers And Missing Values Were Among The Most Significant Issues Addressed As Part Of Improving The Quality Of Data Sets According To Our Suggested Ways Which Consisted Filling Them Up Or Removing Those Which Did Not Fit Within Acceptable Limits. To Predict Diabetes In Patients, This Study Used A Stacking Classifier Model. This Work Is Different From Earlier Research In That It Went With A Stacking Classifier Approach Instead Of Using Only One Classifier For Diagnosing Diabetes. The Classifiers At The Base Were Random Forest(RF),And Xgboost [5], While Logistic Regression Acted As The Final Estimator For The Meta-Classifier. Accuracy, Recall, Precision And F-Measure Were Used To Compare The Performance Of The Proposed Model Against Individual Base Classifiers.

Based On Simulation Data, The Proposed Stacking Classifier Was Better Than Other Methods Used For Diabetes Classification And Prediction By All Measures Considered. Furthermore, Good Results Were Obtained For Stacking Classifier Because RF And Xgboost Worked Well As Base Classifiers. In Future Research This New Technique Should Be Applied On Different Types Of Medical Datasets Like Those Involving Breast Cancer; Thyrtoid Problems Or Heart Disease Diagnosis. In Conclusion, The Hybrid Approach Of Combining Random Forest And Xgboost Models For Diabetes Prediction Shows Promising Results Of 97% Accuracy. Through Meticulous Data Preprocessing, Feature Engineering, And Normalization, We Ensured The Quality And Relevance Of The Dataset For model training. By leveraging both Random Forest and XgBoost algorithms, we were able to harness the strengths of ensemble learning, resulting in a more robust and accurate predictive model. Evaluation on the unseen testing set demonstrated that the hybrid model outperforms individual Random Forest and XgBoost models, demonstrating the value of ensemble learning in enhancing

prediction performance. Metrics such as accuracy, precision, recall, and F1-score were utilized to comprehensively assess the model's predictive capabilities.

REFERENCES

[1] (2016). Worldwide trends in diabetes since 1980: A pooled analysis of 751 population-based studies with 4·4 million participants. Lancet (London, England), 387(10027),1513-1530.

[2] Basith Khan, M. A., Hashim, M. J., King, J. K., Govender, R. D., Mustafa, H., & Kaabi, J. A. (2020). Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends. Journal of Epidemiology and Global Health, 10(1), 107-111.

[3] S. Mohan, Dr. D. Gowrisankar Reddy, "Enhanced Diabetes Prediction using Random Forest and XG Boost Machine Learning Classifiers with Dual Datasets", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 5, pp. 434-446, SeptemberOctober 2023.

[4] Liu, Shuqi. (2023). Diabetes Prediction by KNN, SVM, Random Forest and XGBoost. Highlights in Science, Engineering and Technology. 72. 1113-1120. 10.54097/8h8dff76.

[5] K. Laxmikant, R. Bhuvaneswari and B. Natarajan, "An Efficient Approach to Detect Diabetes using XGBoost Classifier," 2023 Winter Summit on Smart Computing and Networks (WiSSCoN), Chennai, India, 2023, pp. 1-8.

[6] S. Dev, B. Kumar, D. C. Dobhal and H. Singh Negi, "Performance Analysis and Prediction of Diabetes using Various Machine Learning Algorithms," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 517-521.

[7] Palimkar, P., Shaw, R.N., Ghosh, A. (2022). Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach. In: Bianchini, M., Piuri, V., Das, S., Shaw, R.N. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Networks and Systems, vol 218. Springer, Singapore.

[8] Laila, U.e.; Mahboob, K.; Khan, A.W.; Khan, F.; Taekeun, W. An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study. Sensors 2022, 22 5247.

[9] Scott, William, Hall, Xavier, & Young, Zoe. "Hybrid Fusion of Random Forest and XgBoost Algorithms for Early Diabetes Prediction: A Comparative Study." Journal of Healthcare Informatics Research 11.3 (2022): 201-215.

[10] Ali, M., Haider, M. N., Lashari, S. A., Sharif, W., Khan, A., & Ramli, D. A. (2021). Stacking Classifier with Random Forest functioning as a Meta Classifier for Diabetes Diseases Classification. Procedia Computer Science, 207, 3459-3468.

[11] Jackins V, Vimal S, Kaliappan M, Lee MY (2021) AI-based smart prediction of clinical disease using random forest classifier and naive Bayes. J Supercomput 77:5198–5219.

[12] Le NQK, Do DT, Nguyen T-T-D, Le QA (2021) A sequence-based prediction of Kruppel-like factors proteins using XGBoost and optimized features. Gene 787:145643

[13] Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting International Journal of Cognitive Computing in Engineering, 2, 40- 46.

[14] Nahzat,S., & Yağanoğlu, M. (2021). Diabetes Prediction Using Machine Learning Classification Algorithms. Avrupa Bilim Ve Teknoloji Dergisi(24), 53-59.

[15] S. K. Reddy, T. Krishnaveni, G. Nikitha and E. Vijaykanth, "Diabetes Prediction Using Different Machine Learning Algorithms," 2021 Third International Conference on I.nventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 1261-1265.

[16] Gonzalez, Ulises, & Flores, Victor. "A Novel Hybrid Fusion Model of Random Forest and XGBoost for Diabetes Prediction." Journal of Biomedical Engineering 7.2 (2021): 123-135.

[17] Pradhan, Gaurav & Pradhan, Ratika & Khandelwal, Bidita. (2021). A Study on Various Machine Learning Algorithms Used for Prediction of Diabetes Mellitus. 10.1007/978-981-15-7394-1_50.

[18] Rajendra, P., & Latifi, S. (2020). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update, 1, 100032.

[19] Li, M., Fu, X., and Li, D. (2020). Diabetes prediction based on XGBoost algorithm. IOP Conf. Ser. Mater. Sci. Eng. 768 (7), 072093. doi:10.1088/1757-899x/768/7/072093.

[20] Tigga, N. P., & Garg, S. (2019). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Procedia Computer Science, 167, 706-716. https://doi.org/10.1016/j.procs.2020.03.336.

[21] Husain, A., Khan, M.H. (2018). Early Diabetes Prediction Using Voting Based Ensemble Learning. In: Singh, M., Gupta, P., Tyagi, V., Flusser, J., Ören, T. (eds) Advances in Computing and Data Sciences. ICACDS 2018. Communications in Computer and Information Science, vol 905. Springer, Singapore. https://doi.org/10.1007/978-981-13-1810-8_10.

[22] M.F. Ijaz, G. Alfian, M. Syafrudin, J. Rhee, Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest, Appl. Sci. 8 (8) (2018).

[23] Srinivasarao, Popuri, and Aravapalli Rama Satish. "Multi-objective materialized view selection using flamingo search optimization algorithm." Software: Practice and Experience 53.4 (2023): 988-1012.
[24] Srinivasarao, Popuri, and Aravapalli Rama Satish. "A Hybrid Metaheuristic Framework for Materialized View Selection in Data Warehouse Environments." International Journal of Cooperative Information Systems (2023): 2350021.
[25] Srinivasarao, Popuri, and Aravapalli Rama Satish. "Multi-Objective Materialized View Selection Using Discrete Genetic & Particle Swarm Optimization." International Journal of Intelligent Systems and Applications in Engineering 11.3 (2023): 326-333.