Diabetes prediction using Machine Learning Algorithm: A comparative analysis

Viksith Bardia
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Chennai, Tamil Nadu,India
viksithbardia2001@gmail.com

E Sophiya
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Chennai, Tamil Nadu,India
e_sophiya@ch.amrita.edu

2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS) | 979-8-3503-8436-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICACCS60874.2024.10717264

Abstract*—Diabetes prediction is an essential task in healthcare that could be achieved through Machine Learning models. Several factors contribute to diabetes such as overweight, high cholesterol levels or frequent urination. This research includes variety of data sources for predicting diabetes including The RTML dataset consisting data from Bangladeshi women and PIMA Indian dataset. These both combined thereby enriching the dataset's variety and data sources. The mutual information feature selection approach was integrated into this research to find the most relevant features in the dataset. To tackle the issue of class imbalance over-sampling Technique (SMOTE) is utilized to rectify the uneven distribution between diabetic and non-diabetic populations. The classification accuracy of the different Machine Learning(ML) algorithms is compared which include algorithms like random forest, logistic regression, voting classifier, KNN, neural networks, Ada Boost, XgBoost and voting classifier. After testing all the algorithms, the model achieved a cross validation score of 85%, accuracy of 83.22% and AUC of 0.86 using the XgBoost algorithm. The use of explainable AI techniques like the LIME framework helped provide interpretable explanations of the model's predictions.

Keywords: ***Diabetes Prediction, SMOTE, Machine Learning, Accuracy , Ensemble techniques , Interpretability , XG Boost , Cross Validation.***

I. Introduction

According to the researches by Immune Deficiency Foundation(IDF) [1], there is a persistent rise in of diabetes, with the global diabetic population projected to reach around 643 million by 2030. This research paper underscores the paramount importance of diabetes prediction in healthcare, emphasizing its crucial role in combating a global health crisis that affects approximately 537 million individuals. Diabetes ranks as the most widespread and life-threatening non-communicable disease.

Diabetes prediction is essential in healthcare due to its potential to mitigate the burden of this chronic disease through early intervention and personalized management strategies. By leveraging predictive models, healthcare providers can identify individuals at high risk of developing diabetes before clinical symptoms manifest, enabling timely preventive measures such as lifestyle modifications, dietary interventions, and regular monitoring.

Diabetes disrupts the body's ability to effectively process ingested food, leading to an accumulation of excess sugars in the bloodstream. When the pancreas is adversely affected, the insulin generated—essential for carbohydrate digestion loses its effectiveness in breaking down blood sugars derived from food intake. This can manifest as insufficient insulin production or the production of inefficient insulin, resulting in elevated blood sugar levels a condition known as hyperglycemia. The exposure of organs and tissues to heightened sugar levels poses severe risks, cardiovascular ailments as well as potential failures in kidney and liver function. Also, frequent urination can serve as a significant indicator of diabetes due to the body's attempt to eliminate excess glucose through urine, a hallmark characteristic of diabetes. Also, high level of glucose in the bloodstream, overwhelm the kidneys' capacity to reabsorb glucose efficiently.

As per research study [2], diabetes manifests in three primary forms, with type 1 being the most prevalent, affecting a significant number of individuals. In this type, cells fail to produce adequate insulin, leading to compromised immune function. The causation and prevention of type 1 diabetes remain inconclusive. Next the type 2 diabetes is characterized by inadequate insulin production by the body's cells or ineffective utilization of produced insulin. This form is much more predominant affecting approximately 90% of diagnosed diabetes cases.

Gestational diabetes occurs during pregnancy when a woman experiences a sudden increase in blood sugar levels, posing risks to both maternal and fetal health. Interestingly, it often reoccurs in subsequent pregnancies and might start developing type 1 or type 2 diabetes also.

Thereby, the paramount aim of diabetes prediction plays a pivotal role by not only diagnosing patients but also furnishing transparent rationales for the decisions rendered, thereby instilling heightened awareness regarding individuals' prevailing health statuses.

To do this, pinpointing the optimal model for prediction employs a multifaceted methodology and precise Machine Learning models . Some of the major contributions showcased in this research work includes:

Enhancing the ability of the model to make accurate predictions by using two different datasets. The utilization of the Pima Indian diabetes dataset [6] with 769 rows, supplemented by samples from 108 individuals working in a textile factory in Bangladesh [12], reflects a deliberate effort to broaden the dataset's scope and data sources capturing a more comprehen-

979-8-3503-8436-9/24/$31.00©2024 IEEE

sive view of factors influencing diabetes.

The mutual information feature selection technique is utilized to pinpoint the most illuminating attributes within a dataset regarding the target variable, gauging the reciprocal reliance or connection between each attribute and the target variable.

The research paper fortifies its methodology by integrating an extensive spectrum of Machine Learning classification techniques, including logistic regression (LR), K-nearest neighbor (KNN), Decision Tree (DT), AdaBoost, Voting classifier and Extreme Gradient Boosting (XgBoost).The culmination of these efforts is the achievement of an impressive 83.22% accuracy score using the XgBoost classifier.

In this research, the LIME (Local Interpretable Model-agnostic Explanations) framework is integrated to boost the interpretability and transparency of machine learning models utilized in diabetes prediction. LIME strives to furnish localized explanations for the individual forecasts generated by intricate machine learning models.

II. Literature Review

Diabetes significantly impacts a substantial portion of the adult population, prompting numerous research endeavors focused on predicting diabetes symptoms. These studies explore a diverse range of methodologies, encompassing Machine learning (ML), deep learning and adaptive algorithms. In recent times, machine learning algorithms have gained widespread acceptance as a formidable technique, drawing considerable attention from the medical community.

Machine Learning has demonstrated robust predictive capabilities and the ability to analyze numerous variables simultaneously. Additionally, ML has developed techniques for effective variable screening, enabling the identification and comprehension of intricate correlations among variables. Prior research has established the utility of ML as a valuable tool for diabetes prediction.

Zolfaghar [21]conducted a diabetes detection study utilizing a combination of Support Vector Machines Learning algorithm and a fully connected neural network. Following this methodology, the results obtained from the individual classifiers were consolidated through the utilization of the majority voting technique, resulting in superior performance compared to individual classifiers, achieving a success rate of 88.04%.

Edeh et al [8] and colleagues evaluated four distinct algorithms like Naive Bayes, SVM, DT and random forest(RF) on two different diabetes prediction datasets. In the experimentation, the highest accuracy was achieved using SVM, reaching 83.1% with the PIMA dataset.

Dadgar and Kaardaan [9] introduced a novel approach to predict diabetes, combining feature selection using the UTA algorithm with Neural Network (NN). The NN's weights were optimized through the genetic algorithm. This approach achieved a diabetes prediction accuracy of 87.46%.

Transforming Clinical Data into Image Data using deep learning was developed by Muhammet Fatih Aslan [14] .Here the Numerical data is converted into images based on importance features, aiming to leverage Convolutional Neural Network (CNN) models for early diabetes diagnosis. After fine tuning of the ResNet18 model it achieved the accuracy of 80.86% and the ResNet50 accuracy was 80.47% .The SVM algorithm along with the cubic kernel function achieved the highest accuracy of 92.19%. A framework for genetic programming-based diabetes prediction was proposed, and it performed better than other methods they had used.

Hasan et al [6] ,on the other hand explored DT, RF, and Forward Neural Network models alongside with dimensionality reduction techniques such as principal component analysis and redundancy maximum relevance (mRMR) for diabetes prediction. In their study, RF outperformed the other models, boasting a prediction accuracy of 77.21%. Using the PIMA dataset and various ensemble-based machine learning methods, the authors attempted to predict diabetes. The ROC curve metric was employed. Ultimately, the suggested ensemble classifier achieved an AUC of 0.95.

Karthika et al [16] purposed SVM and (RF) ML algorithms to forecast the likelihood of developing diabetes-related diseases. Through meticulous data preprocessing and feature selection techniques like step forward and backward feature selection, relevant features were identified. Leveraging Principle Component Analysis (PCA) for dimensionality reduction, the prediction accuracy reaches 83% with RF, outperforming SVM with an accuracy of 81.4%.

In the study on the prediction of diabetes patients, Maniruzzaman et al [10] utilized four classifiers, namely Random Forest, Naive Bayers, Ada Boost, and Decision tree. The determination of risk factors for diabetic conditions was carried out using Logistic Regression, considering the odds ratio. The employed partitioning techniques were denoted as K2, K5, and K10. The assessment of these classifiers centered on metrics such as accuracy and the area under the curve (AUC). The findings revealed that age, BMI and blood pressure emerged as noteworthy factors associated with diabetes .

The model devised by Nitesh et al [15] utilizes artificial neural networks (ANN) for diabetes detection, showcasing effectiveness in forecasting the survival probability of individuals with diabetes. Utilizing the "Pima Indian Diabetes" dataset, consisting of medical records from 768 patients, the proposed model achieves an impressive accuracy of 85.09%, affirming its efficacy in diabetes prediction.

Zhang et al [20] applied Machine Learning(ML) techniques to conduct a comprehensive analysis for predicting diabetes. Utilizing the advanced Prediction Model Risk of Bias Assessment Tool, the ML models underwent scrutiny to evaluate the potential for bias. The Meta-DiSc software tool was utilized for performing the analysis and assessing variability. The results indicated that machine learning models exhibited enhanced performance in contrast to traditional screening methods for predicting diabetes.

Swapna et al [18] implemented a deep learning methodology for diabetes identification, where the research integrated long short-term memory (LSTM), convolutional neural

network (CNN), and a fused model for capturing dynamic features. These dynamic features were then input into a Support Vector Machine (SVM) for the classification task. The dataset pertaining to heart rate variability (HRV) was employed for diabetes diagnosis through the deep learning approach. The researchers claimed that their system enhances diabetes detection via ECG signals, achieving a heightened accuracy rate of 95.7%.

Sridar et al [17] proposed a medical diagnostic system aimed at predicting diabetes utilizing the backpropagation and Association Rule Mining algorithm. The primary objective of this research was to assess a patient's diabetes risk independently of medical professionals. Clinical data, based on attributes sourced from the Pima Diabetes dataset, was collected during the study. The system incorporated real-time inputs from a glucometer, with certain attributes entered manually. To diagnose diabetes, both the Apriori algorithm and backpropagation algorithm were employed, categorizing patients into three risk classes. The study reported accuracies of 83.5%, 71.2%, and 91.2% from the backpropagation algorithm, Apriori algorithm, and the combined use of both algorithms, respectively.

The primary algorithm investigated in [13] involved Logistic Regression (LR), complemented by various other Machine Learning (ML) approaches such as DT, Naive Bayes, SVM, and KNN to assess performance enhancement. The experiment utilized two main datasets. The initial dataset i.e., Pima Indians, comprised nine features, while the second dataset, Vanderbilt, featured 16 distinct attributes. The outcomes of the study highlighted LR as one of the most effective algorithms for constructing prediction models.

Moreover, in the work by Palimkar et al. [11], another comprehensive model was proposed to enhance the accuracy of diabetes anticipation and classification. The study employed a range of ML algorithms, including RF, AdaBoost, Gaussian Naive Bayes (GNB) and Gaussian process classifier (GPC). The evaluation of these models was based on accuracy, F1 score and error metrics.

III. Proposed System

The core objective of the system is to forecast the probability of diabetes in individuals by considering diverse health related features. This enables timely intervention and personalized healthcare strategies. In doing so, the system aims to provide proactive health insights and enhance individual well-being. Data processing stands as a pivotal step in the proposed pipeline given that higher data quality empowers classifiers to learn more effectively. The objective is to train the model to accurately ascertain whether an individual has diabetes or not.

A. Data Collection

The Pima Indian dataset, collected from the Pima Native American population in Arizona [1], USA, comprises 768 instances and 9 attributes and The RTML dataset comprises data collected from female employees of Rownak Textile Mills Ltd, located in Dhaka, Bangladesh. [12], referred to as the 'RTML dataset' which consists of 108 instances and 7 attributes is used. Table 1 shows the dataset attributes of the combined datasets with their mean and maximum values.

TABLE I
ATTRIBUTES OF THE COMBINED DATASET.

S.NO	Attribute	Average (%)	Maximum
1	Blood Glucose	119.14	239.4
2	Body Mass Index	31.05	67.1
3	Age	33.0	81
4	Blood Pressure	70.39	122
5	Pregnancy	3.54	17
6	Insulin	85.78	846.0
7	Skin Thickness	19.32	99.0

The combination of the dataset results in increased sample size and also capturing a broader range of diversity in terms of geographic locations. These characteristics include vital health and demographic elements, such as pregnancy count, skin thickness, insulin quantities, BMI, blood pressure and age.

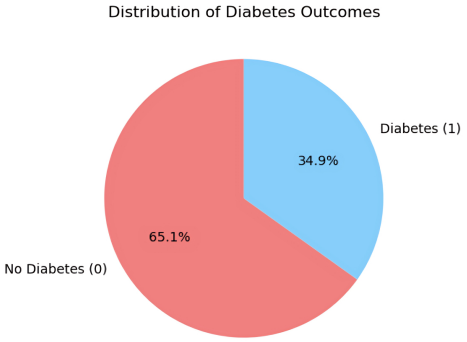

Fig. 1. Percentage of Diabetic and Non diabetic in the combined dataset.

The dataset primarily serves the purpose of supporting diabetes prediction through binary classification, distinguishing individuals into either diabetic (1) or non-diabetic (0).There are 877 rows and 7 attributes as shown in Figure 1.These depict the percentage of diabetic and non-diabetic patients. The following attributes of the dataset are mentioned below:

1) Age: Representing the age of the individual.

2) Glucose: The amount of glucose present in blood, indicated in mg/dL .

3) Insulin: Revealing the insulin concentration in the person's blood measured in µ/ml.

4) Blood Pressure: The pressure of blood pushing against the walls of your arteries, denoted in mm Hg.

5) Pregnancy: Representing the total number of pregnancies pf the woman has experienced.

6) Skin Thickness: Expressing the measurement of the skinfold thickness at the triceps in mm.

7) Body Mass Index (BMI): Metric that utilizes your height and weight to determine whether your body weight is within a healthy range, expressed in kg/m².

8) Outcome: Variable that indicates whether or not diabetes is present in the patient.

B. Data Preparation

In our proposed method, the pre-processing step includes outlier rejection, missing data handling, standardization and addressing class imbalance of the attributes are briefly described as follows:

1) Outlier Rejection: An outlier represents a significantly deviating data point compared to others within the dataset. It is necessary to remove these outliers from the data distribution as classifiers can be highly influenced by the data reach and attribute distribution. The equation for outlier elimination can be expressed as in Equation (1):

$$P(x) = \begin{cases} x, & \text{if } z_1 - 1.5 \times IQR \leq x \leq z_3 + 1.5 \times IQR \\ !x, & \text{otherwise} \end{cases} \tag{1}$$

Given a feature vector x that lies in n-dimensional space ($x \in \mathbb{R}^n$), where x is an instance of the feature vector . $z1$, $z3$ represent the lower quartile and upper quartile. After removal of outliers, the attributes underwent a process to address missing or null values.

2) Filling Missing Values: In our proposed approach, we chose to impute these missing values with the mean of the particular attributes, rather than simply discarding them. This imputation with the mean is valuable technique because it aids in filling gaps in continuous data without introducing new outliers. The mathematical representation of imputing the mean in Equation(2) is as follows:

$$N(x) = \begin{cases} \text{mean}(x), & \text{if } x = 0 \\ x, & \text{if } x! = 0 \end{cases} \tag{2}$$

C. Data Splitting

During the preprocessing stage of the classification task, the dataset undergoes partitioning into training and testing sets. Here, 80% of the data is designated for model training and 20% is reserved for evaluation. Figure 2 illustrates the mutual information feature selection approach measures how much information the presence or absence of a feature contributes to diabetes prediction.

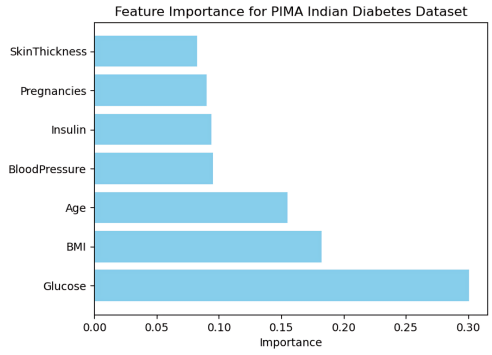

Fig. 2. Importance of each attribute in the dataset.

To tackle the issue of class imbalance and improve model performance, the SMOTE technique was employed to increase the instances of the minority class. In diabetes dataset, the number of individuals without diabetes (majority class) often significantly outweighs the number of individuals with diabetes (minority class). This class imbalance can lead to biased models that perform poorly in predicting the minority class. SMOTE effectively increases the representation of the minority class in the dataset, helping the machine learning model to learn from a more balanced set of examples. Figure 3 depicts the increase of sample from the minority class(diabetes class).

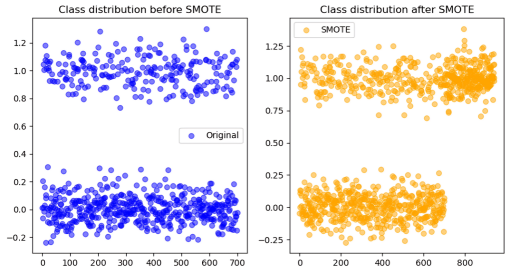

Fig. 3. Oversampling of minority class.

Subsequently, the data underwent standardization which is the process of rescaling of the features in the dataset. This transformation is typically applied to individual features, making them comparable and facilitating better performance for certain Machine Learning algorithms. Standardization is a crucial step in preprocessing as it helps normalize the scales of different features, preventing certain features from disproportionately influencing the model due to their original measurement units. Equation (3) helps in standardization.

$$z_i = \frac{x_i - \mu}{\sigma} \tag{3}$$

The combination of data splitting, SMOTE and standardization lays a solid foundation for the subsequent training and evaluation of the classification model, promoting more robust and unbiased performance metrics.

D. Machine Learning Algorithms

To develop an efficient diabetes prediction system, this study employed various ML algorithms . To prevent overfitting, cross validation and regularization methods has been utilized to fine tune all the ML algorithms.

1) Decision Tree: It is a hierarchical structure that systematically organizes and processes data by making decisions based on the values of features, ultimately leading to predictions or classifications at the leaf nodes. Through hyperparameter tuning we consider a maximum depth of 2 and a minimum sample leaf of 5 using the Gini function as in Equation (4) for classification. [4]

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{4}$$

2) KNN: It is an instance-based learning algorithm, making predictions based on the instances (data points) in the training set. Given a data point, KNN identifies its neighbors which are close to k value in the feature space using the Minkowski distance metric as in Equation(5) with L2 norm, here five nearest neighbors in the feature space are taken to make a classification decision. [6]

$$D = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} \tag{5}$$

3) Random Forest : It is an ensemble model consisting of numerous decision trees, where each tree is built independently using a random subset of the training data Additionally, at every split within each decision tree, a random subset of features is considered. During the training of each tree, a bootstrapped sample (random sample with replacement) is drawn from the original dataset, introducing diversity among the trees. The Gini function is used as the criterion, and the number of decision trees is taken as 8. [11]

4) Logistic Regression: It involves modeling the likelihood of a categorical result based on an input variable. It is particularly suitable for scenarios where the dependent variable is categorical and has two possible classes. This regression technique utilizes the sigmoid (logistic) function equation(6) to convert a linear combination of input features into values ranging between 0 and 1. [13]

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

5) Naive Bayes: A probabilistic Machine Learning algorithm that models the probability of a hypothesis given observed evidence. It is built upon Bayes theorem (7), assuming that class label, the features are independent of each other under the specified conditions. The algorithm leverages this foundation to make predictions efficiently. [7]

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{7}$$

6) AdaBoost: An ensemble learning algorithm that combines predictions from multiple weak learners, often decision trees, to construct a robust classifier. AdaBoost assigns distinct weights to each instance in the dataset, aiming to emphasize misclassified instances during subsequent rounds of training. In this work, the algorithm is specified with 50 estimators and a learning rate of 1.0. [5]

7) XgBoost: Ensemble technique that employs a boosting framework to combine multiple decision trees with the aim of prioritizing error reduction. XgBoost is widely favored for structured data, attributed to its advanced regularization methods and computational efficiency. The learning rate for this research was taken as 0.05 with a maximum tree depth of 5, a minimum child weight of 3, gamma value of 0.1, and column subsampling ratio of 0.3. [19]

8) Voting Classifier: Also, an ensemble method that combines multiple machine learning models and predicts the class label based on the majority vote. In 'hard' voting, each classifier in the ensemble gets one vote, and the class label that receives the majority of votes is predicted. Here A Voting Classifier is created, which combines the predictions from Random Forest, Adaboost, and SVM using a majority voting strategy. Optional hyperparameter tuning is performed using GridSearchCV. [3]

IV. Results and Discussions

In this section, we showcase the outcomes and deliberations pertaining to the suggested automated diabetes prediction system. To gauge the effectiveness of diverse Machine Learning methodologies, we utilized various assessment criteria, encompassing precision (P), recall (R), F1 score, area under curve (AUC) and classification accuracy. These criteria are elucidated by the following equations:

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

Here ,True Positive (TP) corresponds to instances where the model accurately forecasts a positive outcome, False Positive(FP) designates situations in which the model erroneously anticipates a positive result when the actual outcome is negative, True Negative(TN) stands for precise negative predictions and False Negative (FN) takes into account instances where the model inaccurately forecasts a negative result when the actual outcome is positive.

TABLE II
Performance analysis of various algorithms

Classifier	Accuracy (%)	Precision	Recall	F1 score
Decision Tree	73.87	0.63	0.59	0.61
Logistic Regression	75.6	0.64	0.69	0.66
KNN	68.18	0.52	0.75	0.62
Voting Classifier	77.34	0.59	0.67	0.62
Ada Boost	78.4	0.68	0.70	0.69
XG Boost	83.22	0.85	0.88	0.88
Random Forest	80.68	0.75	0.66	0.70
Naive Bayes	76.7	0.65	0.70	0.68

Throughout our investigation, we employed a holdout validation approach, utilizing an 80-20 train-test split with stratification. A comparative analysis of performance metrics for various classifiers using the combined dataset alongside the class balancing (SMOTE) technique is presented in Table 2 for a detailed understanding, Figure 4 illustrates the confusion matrix using XgBoost algorithm.

XgBoost, a powerful gradient boosting algorithm, plays a crucial role in diabetes prediction by offering various capabilities. It can identify significant features for predicting diabetes risk, handle non-linear relationships, and be utilized

as a standalone model or within ensemble learning frameworks to enhance predictive performance. Through hyperparameter tuning, XgBoost can optimize model parameters for improved accuracy.

This visualization highlights that XgBoost effectively classified 125 instances, with 28 true positives and 97 true negatives. Furthermore, the ROC curve value for XgBoost with the SMOTE approach was determined, revealing an AUC value of 0.86 as shown in Figure 5, along with an accuracy rate of 83.22% .

In addition to accuracy, several other metrics were considered for evaluating the classification models. Precision, recall and F1-score for the prediction model were found to be 0.85, 0.88, and 0.87 respectively, offering a thorough evaluation of the model's effectiveness. As an additional metric for evaluating Machine Learning classification models, the cross-validation score was found to be 85%.

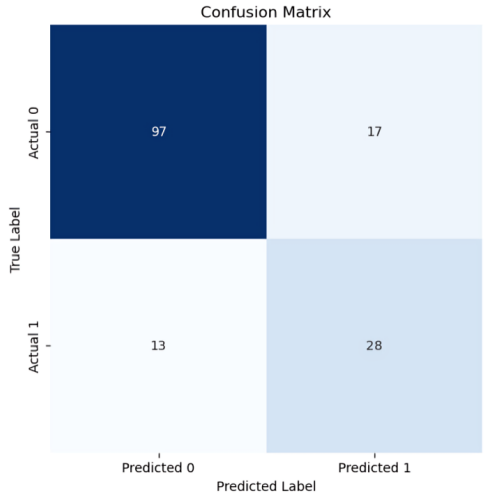

Fig. 4. Confusion Matrix.

To enhance the interpretability of our prediction model on the dataset, we employed the LIME framework , a powerful tool for providing local interpretable explanations of individual predictions. It illustrates that LIME operates under the premise of model-agnostic interpretability, allowing us to understand why the model made specific predictions for individual data points. This is particularly valuable when dealing with complex machine learning models. LIME creates a simplified surrogate model in the vicinity of a particular data point, making it more comprehensible and accessible for human interpretation. In our study, LIME's explanations shed light on the key features allowing us to understand which factors, such as glucose levels, blood pressure or BMI had the most significant impact on a particular prediction.

Table 3 provides a comprehensive comparison of accuracy metrics between the findings of the current research paper and those reported in previous studies. By surpassing the accuracy benchmarks established by previous studies, our research underscores the robustness and efficacy of our predictive modeling techniques in the domain of diabetes prediction.

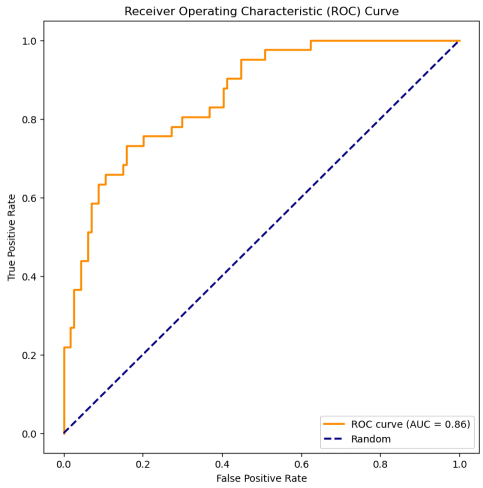

Fig. 5. Area Under Curve (AUC) .

TABLE III
COMPARSION OF ACCURACY WITH PREVIOUS RESEARCH

Author	Classifier	Prior(%)	Proposed(%)
Pranto el al	Random Forest	77.9	80.68
Pranto el al	Decision Tree	72.0	73.87
Jackins et al	Logistic Regression	74.89	77.6
Hosam et al	Voting Classifier	75.0	77.34
Tasin et al	XG Boost	81.0	83.22
Hosam et al	Ada Boost	73.0	78.4
Hosam et al	KNN	64.0	68.18
Jackins et al	Naive Bayes	74.12	76.7

V. CONCLUSION

In conclusion, our research endeavors have delved extensively into the critical challenge of diabetes prediction, recognizing its global impact and the imperative need for robust and precise machine learning models to facilitate early detection and intervention. Through the integration of diverse datasets, mitigation of class imbalance, and application of advanced feature evaluation techniques, our study has underscored the pivotal role of the XgBoost algorithm in classification tasks. We conducted a comprehensive assessment of various machine learning classification methods, demonstrating the efficacy of XgBoost. Additionally, our exploration extended to using interpretable AI methods like the LIME framework. These efforts have yielded valuable insights into the factors influencing predictions, thereby augmenting our understanding of the model's decision-making process. As we progress in understanding diabetes detection, our results highlight the importance of machine learning as a crucial tool in addressing this widespread and impactful disease.

REFERENCES

[1] Diabetes Atlas et al. International diabetes federation. *IDF Diabetes Atlas, 7th edn. Brussels, Belgium: International Diabetes Federation*, 33(2), 2015.

[2] Emma Barron, Chirag Bakhai, Partha Kar, Andy Weaver, Dominique Bradley, Hassan Ismail, Peter Knighton, Naomi Holman, Kamlesh Khunti, Naveed Sattar, et al. Associations of type 1 and type 2 diabetes with covid-19-related mortality in england: a whole-population study. *The lancet Diabetes & endocrinology*, 8(10):813–822, 2020.

[3] Henock M Deberneh and Intaek Kim. Prediction of type 2 diabetes based on machine learning algorithm. *International journal of environmental research and public health*, 18(6):3317, 2021.

[4] Aurélien Géron. Hands-on machine learning with scikit-learn and tensorflow: concepts, tools. *and Techniques to Build Intelligent Systems. nd*, 2017.

[5] Mariwan Ahmed Hama Saeed. Diabetes type 2 classification using machine learning algorithms with up-sampling technique. *Journal of Electrical Systems and Information Technology*, 10(1):1–10, 2023.

[6] Md Kamrul Hasan, Md Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan. Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, 8:76516–76531, 2020.

[7] V Jackins, S Vimal, Madasamy Kaliappan, and Mi Young Lee. Ai-based smart prediction of clinical disease using random forest classifier and naive bayes. *The Journal of Supercomputing*, 77:5198–5219, 2021.

[8] Joshua J Joseph, Prakash Deedwania, Tushar Acharya, David Aguilar, Deepak L Bhatt, Deborah A Chyun, Katherine E Di Palo, Sherita H Golden, Laurence S Sperling, American Heart Association Diabetes Committee of the Council on Lifestyle, Thrombosis Cardiometabolic Health; Council on Arteriosclerosis, Vascular Biology; Council on Clinical Cardiology;, and Council on Hypertension. Comprehensive management of cardiovascular risk factors for adults with type 2 diabetes: a scientific statement from the american heart association. *Circulation*, 145(9):e722–e759, 2022.

[9] Harleen Kaur and Vinita Kumari. Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*, 18(1/2):90–100, 2020.

[10] Md Maniruzzaman, Md Jahanur Rahman, Benojir Ahammed, and Md Menhazul Abedin. Classification and prediction of diabetes disease using machine learning paradigm. *Health information science and systems*, 8:1–14, 2020.

[11] Prajyot Palimkar, Rabindra Nath Shaw, and Ankush Ghosh. Machine learning technique to prognosis diabetes disease: Random forest classifier approach. In *Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2021*, pages 219–244. Springer, 2022.

[12] Badiuzzaman Pranto, Sk Maliha Mehnaz, Esha Bintee Mahid, Imran Mahmud Sadman, Ahsanur Rahman, and Sifat Momen. Evaluating machine learning methods for predicting diabetes among female patients in bangladesh. *Information*, 11(8):374, 2020.

[13] Priyanka Rajendra and Shahram Latifi. Prediction of diabetes using logistic regression and ensemble techniques. *Computer Methods and Programs in Biomedicine Update*, 1:100032, 2021.

[14] Pouya Saeedi, Inga Petersohn, Paraskevi Salpea, Belma Malanda, Suvi Karuranga, Nigel Unwin, Stephen Colagiuri, Leonor Guariguata, Ayesha A Motala, Katherine Ogurtsova, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas. *Diabetes research and clinical practice*, 157:107843, 2019.

[15] P Santhi and S Lavanya. Prediction of diabetes using neural networks. *International Journal of Advanced Science and Technology*, 29(7):1160–1168, 2020.

[16] S Sivaranjani, S Ananya, J Aravinth, and R Karthika. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 141–146. IEEE, 2021.

[17] K Sridar and D Shanthi. Medical diagnosis system for the diabetes mellitus by using back propagation-apriori algorithms. *Journal of Theoretical & Applied Information Technology*, 68(1), 2014.

[18] Goutham Swapna, Soman Kp, and Ravi Vinayakumar. Automated detection of diabetes using cnn and cnn-lstm network and heart rate signals. *Procedia computer science*, 132:1253–1262, 2018.

[19] Isfafuzzaman Tasin, Tansin Ullah Nabil, Sanjida Islam, and Riasat Khan. Diabetes prediction using machine learning and explainable ai techniques. *Healthcare Technology Letters*, 10(1-2):1–10, 2023.

[20] Zheqing Zhang, Luqian Yang, Wentao Han, Yaoyu Wu, Linhui Zhang, Chun Gao, Kui Jiang, Yun Liu, and Huiqun Wu. Machine learning prediction models for gestational diabetes mellitus: meta-analysis. *Journal of medical Internet research*, 24(3):e26634, 2022.

[21] Rahmat Zolfaghari. Diagnosis of diabetes in female population of pima indian heritage with ensemble of bp neural network and svm. *Int. J. Comput. Eng. Manag*, 15(4):2230–7893, 2012.