Comparison of Machine Learning Models for Early Prediction of Diabetes with LIME Interpretability

Rakib Hossain Rifat
Dept. of CSE
BRAC University
Dhaka, Bangladesh
rakib.hossain.rifat@g.bracu.ac.bd
ID: 22366030

Marufa Kamal
Dept. of CSE
BRAC University
Dhaka, Bangladesh
marufa.kamal1@g.bracu.ac.bd
ID: 22366033

Abanti Chakraborty Shruti
Dept. of CSE
BRAC University
Dhaka, Bangladesh
abanti.chakraborty.shruti@g.bracu.ac.bd
ID: 22366034

Ehsanur Rahman Rhythm
Dept. of CSE
BRAC University
Dhaka, Bangladesh
ehsanur.rahman.rhythm@g.bracu.ac.bd

Humaion Kabir Mehedi
Dept. of CSE
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

Annajiat Alim Rasel
Dept. of CSE
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—Diabetes is a disease that gradually begins to affect the physical condition of an individual. It is slowly starting to be rising among the majority of people with increasing time. It can be caused due to hereditary or dietary reasons and lead to an impact on different other parts of the body and create an unhealthy and disastrous lifestyle. In order to detect this incurable and fatal disease this study aims to predict the occurrence of diabetes in a person using machine learning algorithms. The study employs ten supervised and unsupervised learning techniques to analyze a large dataset of medical records with three different approaches. The XAI (Explainable Artificial Intelligence) approach LIME is used to interpret the model's predictions and provide explanations for the results. We achieved 99% accuracy and F1-Score with our best-performing models.**

***Index Terms*—Diabetes, Machine Learning, Logistic Regression, Random Forest Classifier, Decision Tree, K-Nearest Neighbor(KNN), XGBoost, Multi layer Perceptron, AdaBoost, LightGBM, LIME, XAI**

I. Introduction

Diabetes is a chronic illness that affects millions of people worldwide and is one of the main causes of death, morbidity, and disability. It is a dreadful condition caused due to shortage of insulin production by the pancreas or inefficient insulin utilization by the body [1]. There can be two types of diabetes: type-1 diabetes can occur due to insufficient insulin and can be seen mostly in children ranging between 5-7 years of age. [2] Type-2 diabetes is present in more than 90% people with this disease [3] and is caused by ineffective use of insulin. This type is mainly seen in people above 25, but recently Type-2 has also been detected among younger people. Aside from the symptoms of Diabetes, it can cause damage to other organs of the human body such as the heart, nerves, eyes, kidneys, liver, and even the nervous system. The immunity of an individual may also get lower due to having Diabetes. Patients are more prone to have cardiovascular diseases such as heart attacks and strokes [8]. Additionally, diabetic retinopathy, a significant cause of blindness, is brought on by cumulative long-term damage to the retina's small blood capillaries [9]. It also puts an effect on the pregnant women and the baby [21], causes kidney failure, and other health issues. Diabetes may be prevalent for four to twelve years before diagnosis, according to research [10]. If diabetes is not adequately controlled, it can have both short and long-term consequences and lead to complications. To manage the disease and avoid any health damage, early diagnosis, and prompt management are essential. According to the International Diabetes Federation (IDF), around 463 million people had diabetes in 2019, 537 million people had diabetes in the world in 2023, and 90 million people in the SEA Region. This number is expected to rise to 155.1 million by 2045 [15]. Thus calculating the risk factors and their severe effects on the quality of life this paper tries to evaluate different machine learning models and their performances to detect early diabetes.

Machine learning techniques have shown great potential in predicting the onset of diabetes using various clinical and demographic features. ML models like Support Vector Machine(SVM), Logistic Regression(LR), Random Forest Classifier(RF), Decision Tree(DT), K-Nearest Neighbor(KNN), XGBoost, Multilayer Perceptron, AdaBoost, LightGBM and an Ensemble model have been evaluated to predict early diabetes present in the patient or not based on the data received. The UCI Machine Learning Repository [17] has been used for training the models and evaluation purposes. For better interoperability of the ML models, we use a merging discipline explainable AI (XAI) that seeks to make machine learning models transparent and understandable so that both clinicians and patients can use them with confidence. We have used the XAI technique specifically LIME(Local Interpretable Model-Agnostic Explanations) to explain the factors that contribute to the prediction.

II. Related Works

There are many types of research that have been done on the early detection of diabetes using machine-learning techniques. In this section, we are highlighting some of them which have made prominent contributions in the area of early diabetes disease detection.

In the year 2021, Abdulhadi and Al-Mousa [4] proposed a research work that predicts the possible presence of diabetes -specifically in females- at an early stage. The work focused on predicting type-2 diabetes and the authors used Pima Indian dataset [5] in their work. They used various ML algorithms such as Logistic Regression, LDA, SVC, Linear SVC, Random Forest(RF), and voting classifier where the RF showed the best accuracy of 82%. Another paper in the same year by Hassan et al. [6] also focused on detecting early diabetes using machine learning classifiers like Logistic Regression, Random Forest, and XgBoost. They used a dataset collected from Khulna Diabetes Center, Khulna, Bangladesh which had 289 samples of type -2 diabetes. Among the classifiers, Logistic Regression performed well and achieved 88% accuracy. In another work [7] authors used A publicly available dataset of 520 people with 16 features with eight machine learning models to detect early-stage diabetes. The Random Forest classifier achieved the highest accuracy of 98.31%. A comparative analysis regarding the same topic is discussed by Refat et al. [11] where Machine learning(XGBoost, RF, Decision Tree, KNN, etc) and Deep learning techniques(ANN, MLP, LSTM) have been used for comparison. They used the UCI dataset with 17 attributes and the XGBoost classifier achieved an accuracy of nearly 100% and the lowest accuracy was recorded using the KNN classifier. Apratim Sadhu, Abhimanyu Jadli [12] also used the UCI dataset to predict early-stage diabetes using 7 machine learning classifiers. They used accuracy, F1 score, and ROC to measure the performance of their models. Here also the RF classifier outperformed others by nearly 98% accuracy. Adding to the list of research done on detecting diabetes in the year 2021, the authors proposed a method of incorporating data mining techniques to predict the disease [13]. They collected their data from the sector of statistics of the Public Health Institute and used that in the WEKA [14] environment. Simple Logistic, MLP, Logistic, Naive Bayes, Bayes Net, SMO, and C4.5 are used as prediction techniques among which the c4.5 decision tree showed better performance with 79% accuracy. Salliah Shafi and Prof. Gufran Ahmad Ansari also did research [16] on predicting early-stage diabetes disease using machine learning techniques where they proposed a framework that estimates the diabetes disease with maximum precision. They have also used the Pima Indian Dataset from the UCI library. The primary aim of this research was to use the WEKA tools to predict the disease. Another paper [18] uses ML models and analyzes why ML models do not show stable results in this area. They have also taken the computational time into account to decide the best-performing model for detecting the disease. The authors collected datasets from two sources: an automatic electronic recording device and paper records. Eight features were recorded in the dataset including Pregnancy, Glucose, Blood pressure, Insulin, BMI, etc. In the year 2020, MINH LE et al. [19] published A novel wrapper–based feature selection for early diabetes prediction. In their proposed model authors have used the Multi-Layer Perceptron and optimized using the Grey Wolf Optimization (GWO) and an Adaptive Particle Swam Optimization (APSO). They successfully reduced the number of required attributes of MLP and achieved better performances than the state-of-art models when compared. To preprocess the dataset which is collected from [20] they used the IQR method. In terms of performance., 96% accuracy for GWO- MLP and 97% for APGWO - MLP was achieved.

From the above discussion, we can see that over time and specifically in the recent time frame many studies incorporating and visualizing the performance of Machine learning models in the area of diabetes detection at an early stage have been made. Most of the works have achieved previously expected satisfactory results using models like Random Forest or decision trees. However, the development of diabetes diagnosis currently is still in the impoverished phase due to the dearth of effective and robust models with explanations, despite the fact that various ML-based solutions have previously been published in numerous research articles. We tend to analyze the performance of these ML models with the help of XAI and discuss what can be done the improvement of performances.

III. Dataset

For our research, we have used the publicly available Early stage diabetes risk prediction dataset of the UC Irvine(UCI) Machine Learning Repository [20]. This dataset has 520 instances with 16 attributes. This information was gathered by direct questionnaires from Sylhet Diabetes Hospital patients in Sylhet, Bangladesh, and was given the go-ahead by a doctor. 15 attributes or features are categorical and 1 among them is labeled as continuous. Some of them are in medical terms such as Polydipsia(extreme thirst), Polyphagia(excessive hunger), Thrush(a form of yeast infection), Blurred vision(loss of clear vision), Paresis(weakness of voluntary movement), Muscle stiffness(Tight muscles), and Alopecia areata(hair loss in the body) The attribute information of the dataset is shown in the table I.

A. Dataset Distribution

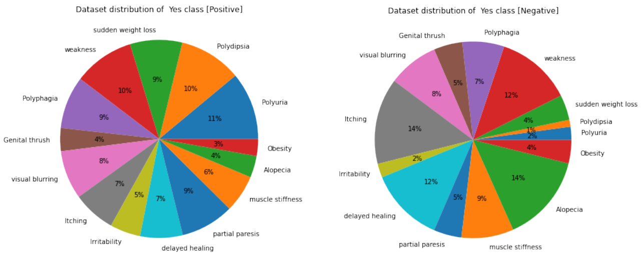
While working with classification-related tasks it is very important to analyze the dataset and visualize the class-wise feature distribution for a better understanding of the dataset.

In our dataset, there are 62% of positive class meaning diabetic class and 38% of healthy class.

From figure 2, we can see the feature-wise distribution for each class and get an overall idea about the impacting features of the classes like from figure 2(a) is seen that in the positive class, the majority of persons are female (around 173) indicating a high risk of diabetes among female individuals.

TABLE I: Dataset Attributes with Example

Attribute Name	Values	Example of The Data
Age	20-65	58
Sex	Male Female	Male
Polyuria	Yes No	No
Polydipsia	Yes No	No
Sudden Weight Loss	Yes No	No
Weakness	Yes No	Yes
Polyphagia	Yes No	No
Genital thrush	Yes No	No
Visual blurring	Yes No	Yes
Itching	Yes No	No
Irritability	Yes No	No
Delayed Healing	Yes No	No
Partial Paresis	Yes No	Yes
Muscle Stiffness	Yes No	No
Alopecia	Yes No	Yes
Obesity	Yes No	No
Class	Positive Negative	Positive



(a) Positive Class

(b) Negative Class

Fig. 1: Impact of Features in Class Distribution

Another interesting aspect is the vast presence of Polydipsia among diabetic patients which is shown in 2(c) around 225 persons who have polydipsia are diabetic. Again from figure 2(d) it can be seen that there are 259 people who are not obese but diabetic. It can indicate that although obesity may not be present in a patient, a person may still be diabetic. The whole feature-wise distribution for a specific class can be found in figure 1.

IV. Methodology

A. *Pre-processing the data*

Since our dataset contains attributes/features that has categorical values such as male, female, yes, no, positive, and negative we have converted these values to numerical data for improving interpretability and simplifying the data representation. We have normalized our dataset to bring all the values to the common range which can improve model performance and ensure convergence.

B. *Correlation of The Features*

After analyzing the dataset we have also tried to understand the correlation between the features of the dataset. Correlation between two variables helps to understand how much the variables are close to one another and can have an effect on the other positively or negatively. We have used Pearson's [23] correlation shown in Figure 3. Based on the correlation, the top 10 features with the highest values are taken to understand the impact of these features on our result. The features selected here are - *Polyuria, Polydipsia, sudden weight loss, partial paresis, Polyphagia, Irritability, visual blurring, weakness, muscle stiffness, Genital thrush, Age, Obesity, delayed healing, Itching, Alopecia, Gender.*

C. *Principal Component Analysis(PCA)*

In order to look further into the top 10 components PCA, or Principal Component Analysis is used as a technique that helps with dimensionality reduction. It operates by locating and eliminating redundant or associated variables from a dataset while preserving the majority of the original variance. Features are selected by choosing the top 10 principal components that account for a certain percentage of the total variance in the dataset. In this case, we have set the threshold percentage to 95%. The features selected here are - *Obesity, Alopecia, Genital thrush, Polyphagia, Irritability, partial paresis, visual blurring, Polydipsia, muscle stiffness, delayed healing*

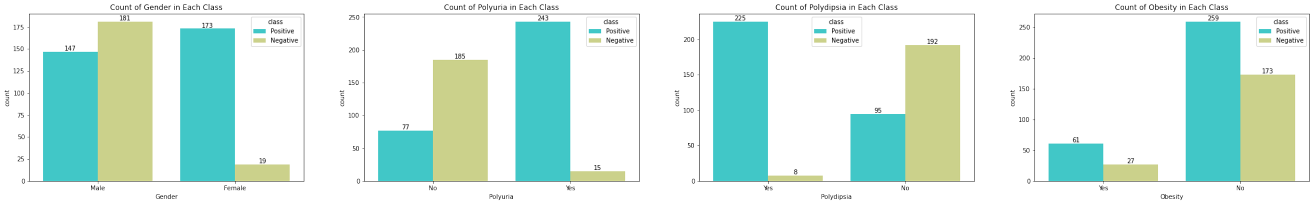
D. *Machine Learning Models*

After finding the top 10 best features through correlation test and PCA along with analyzing all 16 features, we train the model and build a comparison between all the machine learning models in three approaches. The overall process is shown in figure 5. The machine learning models that are used are discussed in brief as follows:

1) Logistic Regression: It [24] is a binary classifier that uses a specified collection of independent factors to predict the categorical dependent variable. It belongs to the category of supervised learning techniques and outputs a binary score based on which we determine whether the patient is diabetic or not.

2) Random Forest Classifier(RF): Random Forest(RF) [25] used for classification and regression task was another algorithm that we have used. RF is quite strong while identifying overfitting and can be quite efficient. Additionally, it can manage missing values in the data and be applied to feature selection. The random state was set to 10 in our model.

3) Support Vector Machine(SVM): Support vector machine(SVM) [26] used for classification tasks tries to locate the hyperplane which maximizes the margin in between the various classes of the data provided. We have used this to understand the results and how our model performed on our



(a) Gender

(b) Polyuria

(c) Polydipsia

(d) Obesity

Fig. 2: Attribute wise Class Distribution

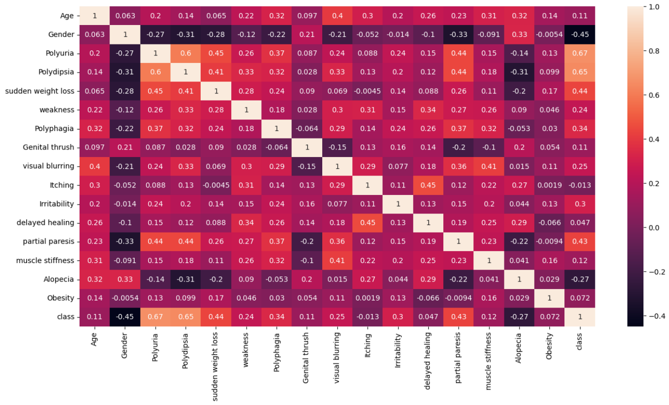


Fig. 3: Correlation between the features of the dataset

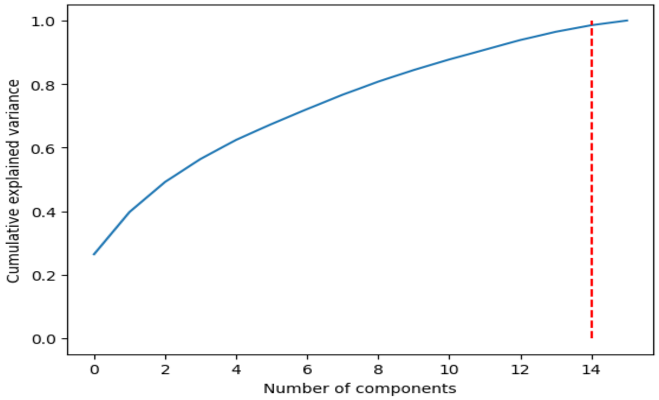


Fig. 4: Reduction of multi-dimensional data to 10 principal components for visualization purposes.

data. The tradeoff parameter C is set to 10 to make the model more strict in terms of minimizing the error and *gamma* value is set to 0.01.

4) Decision Tree(DT): Decision tree [27] is a machine learning technique that builds hierarchical structure by partitioning the data into smaller subsets based on the features. By studying decision rules inferred from the data features, the main goal is to build a tree to predict the target decision. This technique can be beneficial in classification and regression tasks and also handle categorical or numerical data. Our model uses a random state of 1234 and a maximum depth of 3.

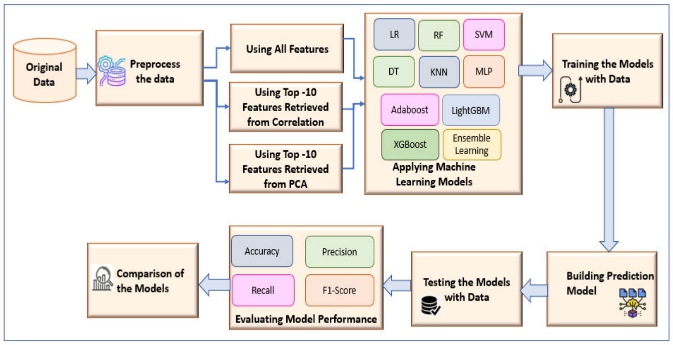

Original Data
Preprocess the data
Using All Features
Using Top -10 Features Retrieved from Correlation
Using Top -10 Features Retrieved from PCA
LR
RF
SVM
DT
KNN
MLP
Adaboost
LightGBM
XGBoost
Ensemble Learning
Applying Machine Learning Models
Training the Models with Data
Building Prediction Model
Testing the Models with Data
Accuracy
Precision
Recall
F1-Score
Evaluating Model Performance
Comparison of the Models

Fig. 5: System Architecture of the Entire Process

5) K - Nearest Neighbor(KNN): KNN [28] is a nonparametric algorithm that tries to make predictions based on the similarity of the data points on which it has been trained and the distance metric provided. When using regression or classification, KNN determines the majority class (classification) or mean value (regression) based on the distances between the query data point and the training set. Finding the value of K can become complex and act as a disadvantage. In our model, 10 neighbors are set and using the Euclidean distance for calculating the distance among the data points.

6) Multi-layer Perceptron(MLP): MLP [29] is a well-known neural network architecture made up of numerous layers of interconnected nodes that process the input data and discover intricate patterns through back-propagation. This process creates a nonlinear function model that makes it possible to predict output data from the input data. Images, text, and time series data can all be used with MLP. *ReLU* activation function, *adam* optimizer and alpha regularization parameter of *0.0001* is set in our MLP model.

7) AdaBoost: For classification tasks, a common ensemble learning technique called AdaBoost (Adaptive Boosting) [30] is utilized. It functions by rating the significance of each sample according to its classification performance and iteratively training weak classifiers on the misclassified examples. The combined weighted votes of the weak classifiers yield the final prediction. AdaBoost can manage noisy data and increase the accuracy of the underlying classifiers. A learning rate of 0.01 and 250 n_estimators was set in our model.

8) LightGBM: Light GBM (Light Gradient Boosting Machine) [22] is an effective gradient boosting system that can handle big datasets with high-dimensional features. It is also made to be effective, scalable, and versatile. To increase this algorithm's precision, Light GBM applies a gradient-based optimization technique with a tree-based model. It also uses a histogram approach to minimize the computation cost while splitting. A learning rate of 0.43 and 500 n_estimators has been used in our classifier.

9) XGBoost: XGBoost [31] or Extreme Gradient Boosting is an ensemble machine learning algorithm. With a tree-based model and regularization to avoid overfitting, it is an extension of the gradient-boosting approach. Its parallel processing and tree pruning techniques make it a well-known classifier for high accuracy prediction and speed. We have used the random state of *42* and set *binary:logistic* to the objective function for our binary classification problem.

10) Ensemble Learning: We have used a voting classifier [32] which is a technique for ensemble learning that combines the predictions of various machine learning models. The goal is to combine many models to increase the reliability and accuracy of the predictions. In our study, we have combined the predictions received from Random Forest, K-Nearest Neighbor(KNN), and LightGBM.

E. Evaluation Matrices

A variety of performance metrics were used in this investigation to explain why ML models could perform well with one evaluation metric's measurement while performing not so great with another metric's assessment. Different evaluation criteria must be used to make sure an ML model is functioning properly and optimally. In this study, we mainly used Accuracy, Precision, Recall, and F1-Score as performance evaluation metrics.

1) Accuracy: Accuracy is defined as the total number of accurate predictions divided by the total number of data samples present in the dataset as shown in the equation (1)-

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2) Precision: The Precision is defined as the total number of accurate positive predictions divided by the total number of positive predictions as shown in the equation (2)-

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

3) Recall: The recall is defined as the total number of accurate positive predictions divided by the total number of actual positive predictions as shown in the equation (3)-

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

4) F1-Score: F1-Score is the harmonic mean of precision and recall as shown in the equation (4)-

$$F1{-}Score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \quad (4)$$

Here,
TP = True Positive , TN = True Negative
FP = False Positive, FN = False Negative

F. LIME Interpretation Model

LIME is an interpretability technique used to explain the predictions made by an ML model. We have used this LIME technique to understand our best and worst-performing models and the impact of the features on determining the results.

V. Experimental Results and Discussion

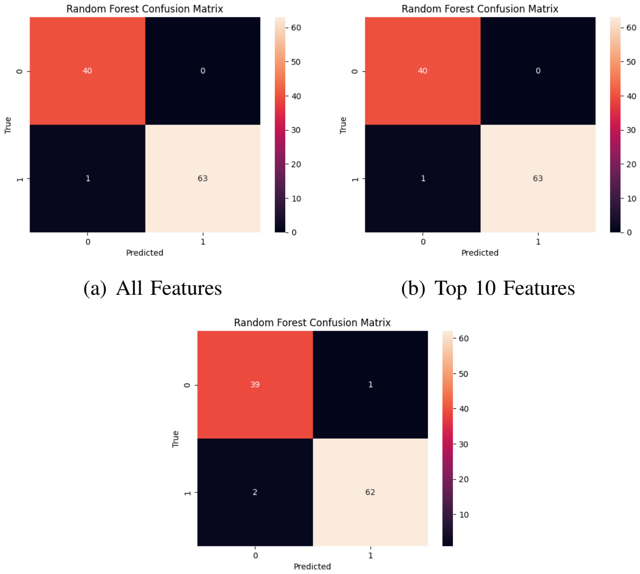
The diabetes risk prediction dataset is used to test each of the aforementioned machine learning models, and the results are stored in terms of accuracy, precision, recall, and f1-score. To detect diabetes in an early stage we used our dataset in two approaches. We wanted to experiment and compare to find out the most effective model for predicting the outcome. First, we used all of the features present in the dataset in our models and predicted the outcome. In a second approach, we used only the top 10 features and used them in our aforementioned models to predict the outcome. The detailed comparison of evaluation metrics results can be found in the table II.

From table II, we can see the performance comparison of all three of our approaches. It is evident that in both approaches 1 and 2 Random Forest classifier outperformed all other classifiers. It achieved equal accuracy, precision, recall, and F1-Score and that is 99% in both of the approaches. However, in approach 3 KNN, LGBM, XGBoost, and Ensemble models outperformed others with 98% of equal performance. Now, if we look into the approach 1 results where all features are used, it can be seen that there are also three more classifiers that achieved the same evaluation scores as Random Forest. The classifiers are KNN, LightGBM, and XGBoost. However, when top-10 features using correlation are only used, the performance of these models decreases a little. KNN achieves an accuracy of 96% and LGBM achieves 97% accuracy compared to the 99% accuracy of approach 1 of both classifiers. But these classifiers still perform better than the other classifiers of the research. The lowest performance is noticed by the Multi-layer perceptron model in approach-1 which is 71% of accuracy and 62% recall and 61% of F1-score. Nevertheless, the interesting factor is the performance of this classifier increases when only top-10 features are used. The accuracy, recall, and F1-score get increased to 95%, 94%, and 95% respectively for approach 2 which is a satisfactory increase. However, the precision of the MLP model is adequate in both of the approaches. All other classifiers like LR, SVC, DT, and AdaBoost also show satisfactory results in all of the approaches with all having accuracy above 90%. Moreover, in approach-3, we can see that the ensemble model performed well along with other models but AdaBoost model performance scores are lower than MLP for this approach. This shows that when features are fewer AdaBoost model performance decreases.

TABLE II: Results on Test Data.
Here, A1 - Using All Features, A2 - Using Top-10 Features from correlation and A3 - Using Top-10 Features from PCA

Model	**Accuracy(%)**			**Precision(%)**			**Recall(%)**			**F1-Score(%)**		
	A1	**A2**	**A3**	**A1**	**A2**	**A3**	**A1**	**A2**	**A3**	**A1**	**A2**	**A3**
Logistic Regression	96	95	88	96	95	87	96	95	88	96	95	87
Random Forest	**99**	**99**	97	**99**	**99**	97	**99**	**99**	97	**99**	**99**	97
SVC	97	96	91	97	96	91	97	96	91	97	96	91
Decision Tree	93	93	87	93	93	86	93	93	86	93	93	86
KNN	**99**	96	**98**	**99**	96	**98**	**99**	96	**98**	**99**	96	**98**
MLP	71	95	83	84	96	84	62	94	86	61	95	83
AdaBoost	90	90	81	91	91	81	89	89	82	90	90	81
LGBM	**99**	97	**98**	**99**	97	**98**	**99**	98	**98**	**99**	97	**98**
XGBoost	**99**	98	**98**	**99**	98	**98**	**99**	98	**98**	**99**	98	**98**
Ensemble	98	98	**98**	98	98	**98**	98	98	**98**	98	98	**98**

From the confusion matrix of the best and worst performing models, we can get an in-depth idea about the strong and weak aspects of the model performance. From figure 6 it can be observed that the RF classifier correctly classifies all of the healthy classes (40) correctly and misclassifies only one diabetic patient to the healthy class in both of the approaches.



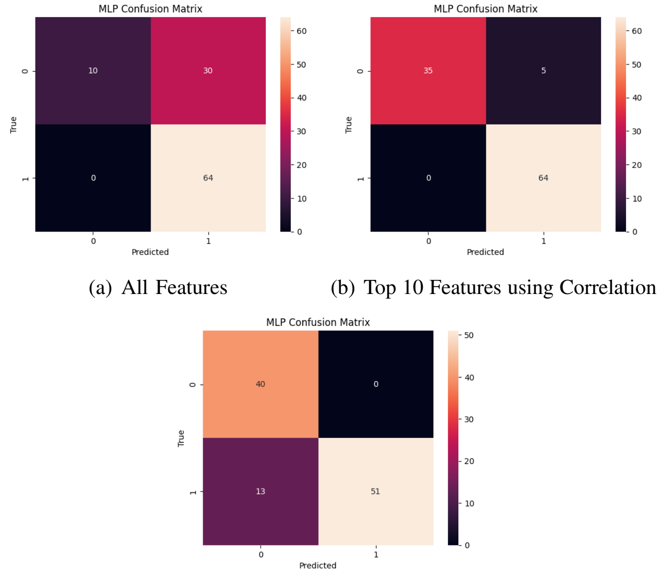
(a) All Features

(b) Top 10 Features

(c) Top 10 Features using PCA

Fig. 6: Random Forest Confusion Matrix

In figure 8, we can see the results for the XGBoost classifier where XGBoost showed the best performance for approach-1 where the model only misclassified one data point. In approach-2 and approach-3 it misclassified two data points. But in approach-2 the model misclassified both diabetic patients as healthy patients. On the other hand, in approach-3, one healthy data and one diabetic data are misclassified. Now coming to the lowest performing classifier MLP shown in figure 7, it accurately predicts all of the diabetic data points but fails to predict 30 healthy data points among 40 data in approach-1. It somehow labels all of the healthy classes as the diabetic class which is the reason for the worst performance of the classifier. In approach-2, the performance improves when only 5 healthy patients are predicted as diabetic patients.



(a) All Features

(b) Top 10 Features using Correlation

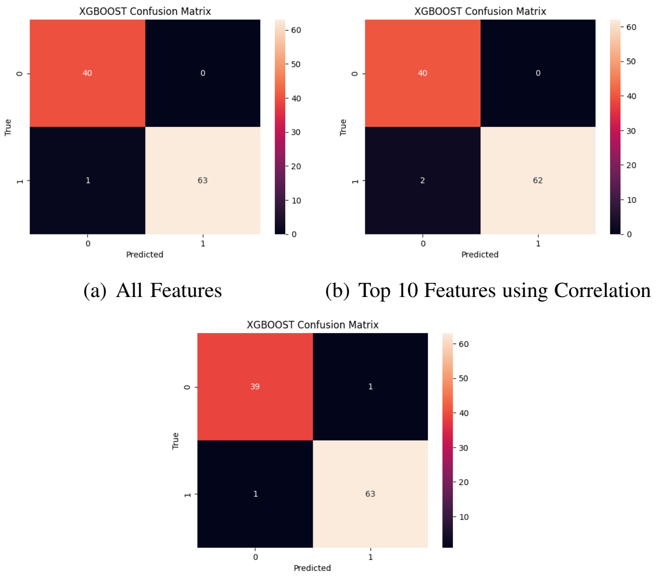
(c) Top 10 Features using PCA

Fig. 7: MLP Confusion Matrix

A. LIME Interpretation of Proposed ML Models

In figure 9, 10 and 11 we can see the LIME interpretations of the best and worst performing models of our three approaches. Here we used LIME to explain and interpret one test data of our dataset and could see the difference in the prediction outcomes and most affecting features.

For approach-1 in figure 9, when all the features are used, we can see that for the best performing Random Forest model in 9(a), the model is 61% sure that the patient is diabetic and Polyuria, Irritability, and genital thrush are the features which affected the decision most. However, for the worst performing model MLP, in 9(b) we can see that along with these three features, Alopecia, Muscle stiffness, and sudden weight loss also impacted the decision, and the model became 100% sure that this data is of a diabetic patient. For approach-2, we

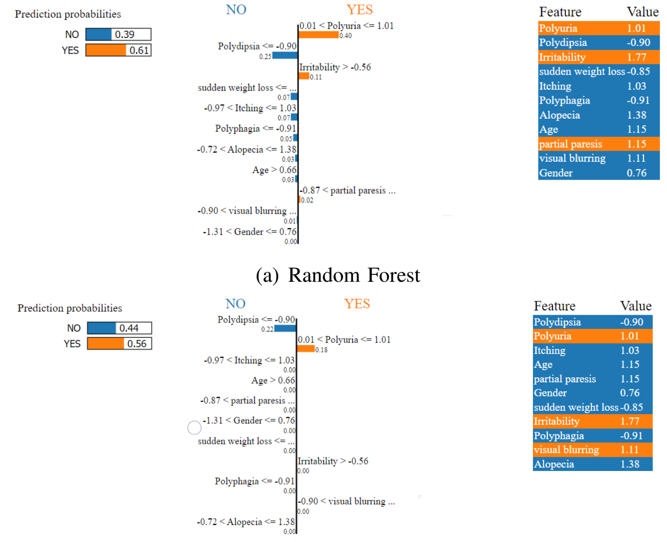


(a) All Features

(b) Top 10 Features using Correlation

(c) Top 10 Features using PCA

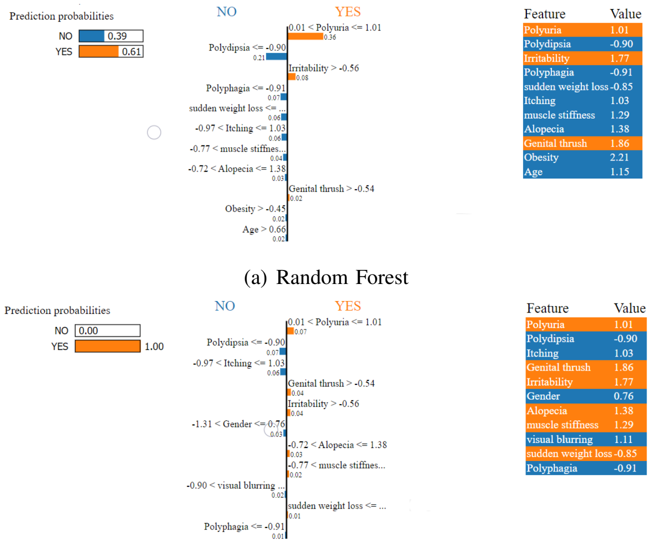
Fig. 8: XGBoost Confusion Matrix



(a) Random Forest

(b) AdaBoost

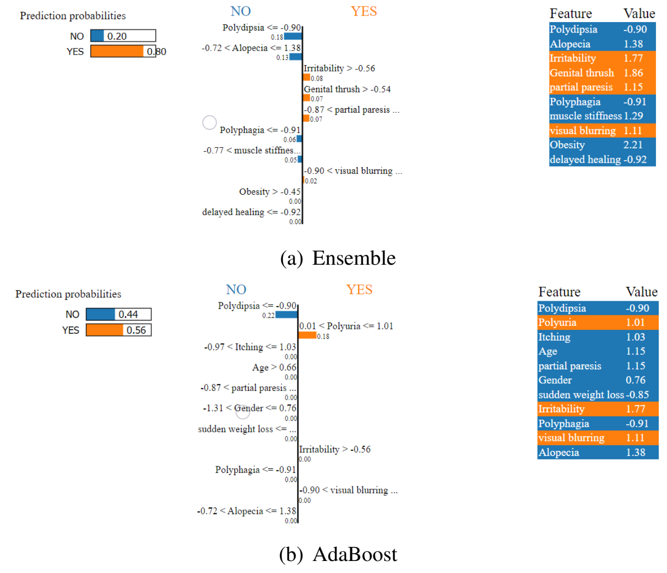
Fig. 10: Approach-2 Lime Interpretations for Best and Worst Performing Model



(a) Random Forest

(b) MLP

Fig. 9: Approach-1 Lime Interpretations for Best and Worst Performing Model



(a) Ensemble

(b) AdaBoost

Fig. 11: Approach-3 Lime Interpretations for Best and Worst Performing Model

can see that in figure 10, Random Forest is again the best-performing model as shown in figure 10(a) which is more confident than the worst-performing AdaBoost model shown in Figure 10(b) where only AdaBoost is mostly impacted by Polyuria and Polydipsia for positive and negative class respectively, on the other hand, Random Forest takes account of other features(Irritability, Sudden weight loss, etc) for the decision. In figure 11 we can see the LIME interpretations for approach-3. Here, the Ensemble model is shown as best performing model and AdaBoost is the worst-performing model again. The feature impact of AdaBoost shown in figure 11(b) is the same as approach-2 here. For the Ensemble model, we can see that in figure 11(a), the model predicts the data as diabetic patients data with 80% confidence. Irritability, Genital Thrush, Partial paresis, and Visual blurring are the highest affecting features in this diabetic class classification while Polydipsia and Alopecia mostly impacted this data will be a healthy class data. We can see that, some of the features are most prominent in the predictions in all of the approaches which include Polyuria, Polydipsia, Genital Thrush, Irritability, Alopecia, etc. This gives us an explanation of why our model predicts the wrong class. By these interpretations, we can also understand which symptoms can be the most important factor in the early detection of diabetes.

In summary, in a field like early diabetes detection, it is crucial that a prediction model is able to identify accurately the diabetic patients compared to the healthy patients. As identifying diabetic patients as healthy patients can cause many more complications and mistreatment in the medical sector. In a field like this, it is essential that the precision and f1-score are taken more into account than just the accuracy. We could see that despite performing poorly with accuracy, the precision is always up to the mark for the MLP model. And of course, all of the classifiers have satisfactory precision and F1-score whereas the best-performing models achieved 99% precision, recall, and F1-score. Also, in disease prediction models it is a must that the predictions are explained and impacting features are identified that's why we tried to understand our best and worst models in all three approaches using LIME.

VI. Conclusion and Future Works

One of the most commendable achievements we have made so far was the use of technology in the medical field. With this use, we could save lives and detect diseases at the early stage to prolong human life span. Detection of early-stage diabetes can ensure appropriate medical care and machine learning models can be used to complete this task with efficiency. Keeping this idea in mind, in this research, we used ten machine-learning models such as - Logistic Regression, Random Forest, SVM, Decision Tree, KNN, MLP, AdaBoost, LGBM, XGBoost, and Ensemble model in three different approaches to predict early-stage diabetes and explained the results with Explainable-AI (LIME). In our experiments, Random Forest, KNN, and XGBoost models performed best with 99% accuracy and F1-Score. We also tried to draw a conclusion about the most affecting symptoms of early diabetes using the LIME in which we could see Polyuria, Polydipsia, Genital Thrush, Irritability, and Alopecia has the most effects. Although we achieved satisfactory performance in all three approaches, we had some limitations like our dataset size was not up to the mark. Also, there was a little class imbalance in the dataset. Some of the additional effecting features could also be incorporated into this research. For future work, we plan to build our own medical dataset of diabetic detection using all the impacting features. We also plan to develop a diabetes risk prediction application using the best model.

References

[1] World Health Organization,"Diabetes", https://www.who.int/news-room/fact-sheets/detail/diabetes, March 25, 2023.

[2] Atkinson, Mark A., George S. Eisenbarth, and Aaron W. Michels. "Type 1 diabetes." The Lancet 383.9911 (2014): 69-82

[3] Chatterjee, Sudesna, Kamlesh Khunti, and Melanie J. Davies. "Type 2 diabetes." The lancet 389.10085 (2017): 2239-2251.

[4] Abdulhadi, N. and Al-Mousa, A., 2021, July. Diabetes detection using machine learning classification methods. In 2021 International Conference on Information Technology (ICIT) (pp. 350-354). IEEE.

[5] L. O. Schulz, P. H. Bennett, E. Ravussin, J. R. Kidd, K. K. Kidd, J. Esparza and M. E. Valencia, "Effects of Traditional and Western Environments on Prevalence of Type 2 Diabetes in Pima Indians in Mexico and the U.S.," Diabetes Care, vol. 29, no. 8, pp. 1866–1871, 2006.

[6] Hassan, M. M., Peya, Z. J., Mollick, S., Billah, M. A. M., Shakil, M. M. H., & Dulla, A. U. (2021, July). Diabetes prediction in healthcare at early stage using machine learning approach. In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 01-05). IEEE

[7] Ergün, Ö. N., & İLHAN, H. O. (2021). Early stage diabetes prediction using machine learning methods. Avrupa Bilim ve Teknoloji Dergisi, (29), 52-57.

[8] Williams, R., L. Van Gaal, and C. Lucioni. "Assessing the impact of complications on the costs of Type II diabetes." Diabetologia 45 (2002): S13-S17.

[9] GBD 2019 Blindness and Vision Impairment Collaborators. and Vision Loss Expert Group of the Global Burden of Disease Study. "Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study." The Lancet. Global health vol. 9,2 (2021): e144-e160.

[10] Harris, Maureen I., et al. "Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis." Diabetes care 15.7 (1992): 815-819.

[11] Refat, M. A. R., Al Amin, M., Kaushal, C., Yeasmin, M. N., & Islam, M. K. (2021, October). A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach. In 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC) (pp. 654-659). IEEE.

[12] Sadhu, A., & Jadli, A. (2021). Early-stage diabetes risk prediction: a comparative analysis of classification algorithms. International Advanced Research Journal in Science, Engineering and Technology (IARJSET), 8(2), 193-201.

[13] Llaha, O., & Rista, A. (2021, May). Prediction and Detection of Diabetes using Machine Learning. In RTA-CSIT (pp. 94-102).

[14] Frank, Eibe & Hall, Mark & Holmes, Geoffrey & Kirkby, Richard & Pfahringer, Bernhard & Witten, Ian & Trigg, Len. (2010). Weka-A Machine Learning Workbench for Data Mining. 10.1007/978-0-387 09823-4_66.

[15] International Diabetes Federation,"Home", https://idf.org/our-network/regions-members/south-east-asia/members/93-bangladesh.html, March 25, 2023.

[16] Shafi, S., & Ansari, G. A. (2021, May). Early prediction of diabetes disease & classification of algorithms using machine learning approach. In Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021).

[17] Dua, Dheeru, and Casey Graff. "UCI machine learning repository, 2017." URL http://archive. ics. uci. edu/ml 7.1 (2017).

[18] Mitushi, Soni, and Varma Sunita. "Diabetes prediction using machine learning techniques." International Journal of Engineering Research & Technology (IJERT) (2020).

[19] Le, T. M., Vo, T. M., Pham, T. N., & Dao, S. V. T. (2020). A novel wrapper–based feature selection for early diabetes prediction enhanced with a metaheuristic. IEEE Access, 9, 7869-7884.

[20] Early stage diabetes risk prediction dataset. (2020). UCI Machine Learning Repository.

[21] Centers for Disease Control and Prevention,"What is diabetes?", https://www.cdc.gov/diabetes/basics/diabetes.html, July 07, 2022.

[22] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).

[23] Freedman, D., Pisani, R., & Purves, R. (2007). Statistics (international student edition). Pisani, R. Purves, 4th Edn. WW Norton &amp; Company, New York.

[24] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–232.

[25] Biau, Gérard, and Erwan Scornet. "A random forest guided tour." Test 25 (2016): 197-227.

[26] Noble, William S. "What is a support vector machine?." Nature biotechnology 24.12 (2006): 1565-1567.

[27] Song, Yan-Yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." Shanghai archives of psychiatry 27.2 (2015): 130.

[28] Guo, Gongde, et al. ”KNN model-based approach in classification.” On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings. Springer Berlin Heidelberg, 2003.

[29] Gardner, Matt W., and S. R. Dorling. ”Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences.” Atmospheric environment 32.14-15 (1998): 2627-2636.

[30] Schapire, Robert E. ”Explaining adaboost.” Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik (2013): 37-52.

[31] Chen, Tianqi, et al. ”Xgboost: extreme gradient boosting.” R package version 0.4-2 1.4 (2015): 1-4.

[32] Dietterich, Thomas G. ”Ensemble methods in machine learning.” Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1. Springer Berlin Heidelberg, 2000.