Early Diabetes Prediction using Machine learning techniques with XAI

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 slowly starts to deteriorate the health of an individual and is seen to be rising among the majority of people with 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 nine supervised and unsupervised learning techniques to analyze a large dataset of medical records. The XAI (Explainable Artificial Intelligence) approaches SHAP and LIME are used to interpret the model's predictions and provide explanations for the results.**

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

I. Introduction

Millions of individuals throughout the world suffer from the chronic disease diabetes, which is a leading cause of mortality, morbidity, and disability. Diabetes is a dreadful condition brought on by either insufficient insulin production by the pancreas or inefficient insulin utilization by the body [1]. There can be two types of diabetes: type-1 diabetes is caused by insulin deficiency and occurs mostly in children 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 was 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]. Moreover, a significant contributor to blindness, diabetic retinopathy results from cumulative long-term harm to the retina's tiny blood vessels [9]. It also puts an effect on the pregnant women and the baby [23], 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 such as Support Vector Machine(SVM), Logistic Regression, Random Forest Classifier, Decision Tree, K-Nearest Neighbor(KNN), XGBoost, Multilayer Perceptron, AdaBoost, LightGBM 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 will use XAI techniques such as Shapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) 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, and voting classifier where the Random forest 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, Random Forest, 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 seven machine learning classifiers. They used accuracy, F1 score, and ROC to measure the performance of their models. Here also the Random Forest 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 on predicting early-stage diabetes disease using machine learning techniques such as SVM, Naive Bayes, and decision trees 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. As performance measurement techniques, they used Certainty, Precision, Recall, F-measure, and ROC. The Naive Bayes classifier got 74.28% accuracy whereas SVM got only 63.10% accuracy. Another paper [] 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. Random Forest classifier here also outperforms others with 80% accuracy. In the year 2020, MINH LE et al. [19] 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. In 2022, Dutta et al. proposed a method of identifying diabetes at an early stage using an ensemble of machine learning models. They introduced a new database on diabetes from Bangladesh. They used a weighted ensemble of NB, RF, DT, XGB, and LGM. They also performed a feature selection and K-fold validation on their model. ANOVA test showed that by ensembling DT + RF + XGB + LGB, the proposed model achieves an accuracy of 73.5%.

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 [22]. This dataset has 520 instances with 16 attributes. This has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh, and approved 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

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

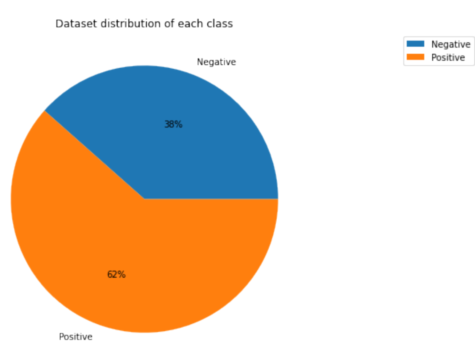


Fig. 1: Dataset Distribution

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 as shown in the figure 1

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 181) indicating a high risk of diabetes among female individuals.

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(p) it can be seen that there are 259 people who are obese but not diabetic. It can indicate that although obesity maybe present in diabetic patient it is not a major factor in diabetes. The whole feature-wise distribution for a specific class can be found in figure 3.

IV. Methodology

A. *Support Vector Classifier*

B. *Random Forest Classifier*

C. *Decision Tree*

D. *Logistic Regression*

E. *K-Nearest Neighbor(KNN)*

F. *AdaBoost*

G. *XGBoost*

H. *Multi-layer Perceptron*

I. *LightGBM*

V. Experimental Results and Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

VI. Conclusion and Future Works

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

References

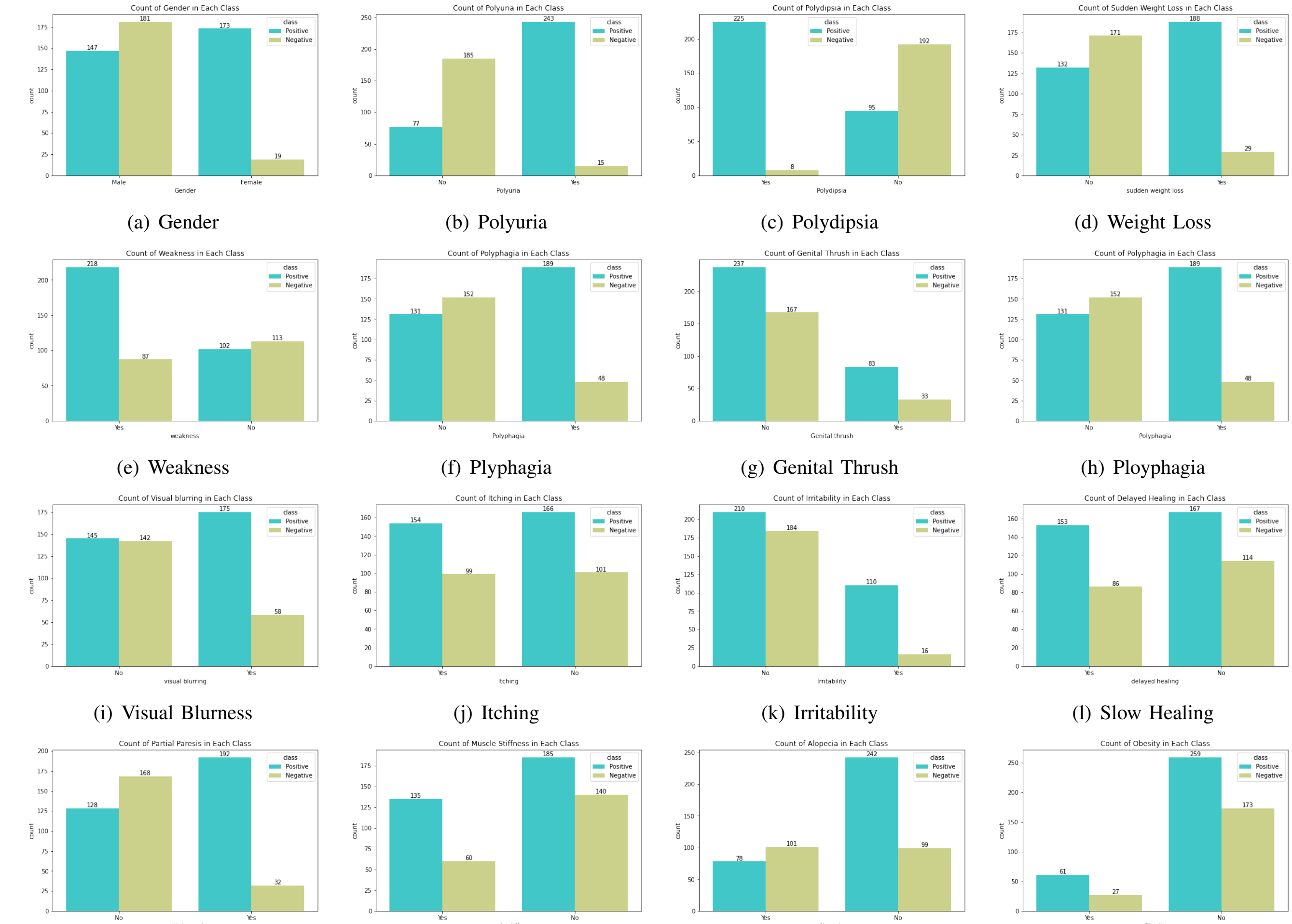
[1] World Health Organization,"Diabetes", https://www.who.int/news-room/fact-sheets/detail/diabetes, March 25, 2023.

[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.



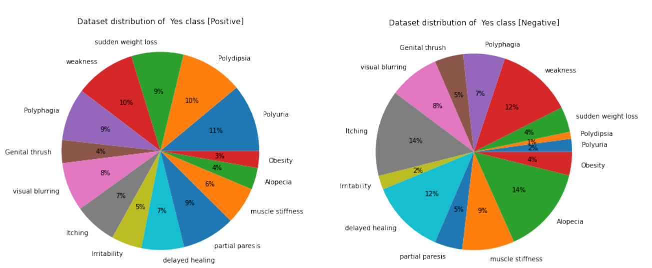
(a) Gender (b) Polyuria (c) Polydipsia (d) Weight Loss

(e) Weakness (f) Plyphagia (g) Genital Thrush (h) Ployphagia

(i) Visual Blurness (j) Itching (k) Irritability (l) Slow Healing

(m) Partial Paresis (n) Muscle Stiffness (o) Alopecia (p) Obesity

Fig. 2: Attribute wise Class Distribution



(a) Positive Class (b) Negative Class

Fig. 3: Impact of Features in Class Distribution

[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] S. K. Somasundaram and P. Alli, "A machine learning ensemble classifier for early prediction of diabetic retinopathy," J. Med. Syst., vol. 41, no. 12, p. 201, Nov. 2017, doi: 10.1007/s10916-017-0853-x

[21] Dutta, A., Hasan, M. K., Ahmad, M., Awal, M. A., Islam, M. A., Masud, M., Meshref, H. (2022). Early prediction of diabetes using an ensemble of machine learning models. International Journal of Environmental Research and Public Health, 19(19), 12378.

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

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