Early-Stage Diabetes Risk Prediction Using Supervised Machine Learning Algorithms

Taminul Islam
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
taminul@ieee.org

Md Rezwane Sadik
Department of Economics & Decision Sciences
University of South Dakota
Vermillion, SD, United States
mdrezwane.sadik@coyotes.usd.edu

Md. Fajle Rabbi Islam
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
fjrabbi.stu@gmail.com

Tanzina Rahman Mona
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
tanzina15-2812@diu.edu.bd

Tanjila Rahman
Department of Computer Science
American International University Of Bangladesh
Dhaka, Bangladesh
tanjilarahman8265452@gmail.com

Md. Musfiqur Rahman Foysal
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
musfiqurrahmanfoysal77@gmail.com

2023 2nd International Conference on Futuristic Technologies (INCOFT) | 979-8-3503-0884-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/INCOFT60753.2023.10425305

***Abstract*— Diabetes is a common health problem worldwide; it is especially pervasive in Bangladesh. The condition manifests in a person when his blood sugar is consistently high. It also contributes to other health problems like blindness, renal failure, heart attack, and stroke. If you know about the early stage, you can take charge and maybe save someone's life. Sadly, this illness is spreading rapidly. The purpose of this research was to quantitatively evaluate the effectiveness of many widely used Machine Learning methods. The medical field is only one area that has benefited greatly from recent advancements in Machine Learning technology. Machine learning algorithms come in a wide variety. Nevertheless, in this research we employ five well-known machine learning algorithms to determine performance metrics: Gaussian Naive Bayes, Random Forest, Support Vector Machine, Logistic Regression, and the Decision Tree classifier. Using real data from diabetic patients in Bangladesh, these algorithms were developed and evaluated. There are 3837 patient records in the dataset, 3057 of which correspond to affected cases and 396 were normal. Out of 5 different machine learning algorithms, Random Forest achieved the highest 98% accuracy.**

Keywords*— diabetics prediction, medical disease prediction, supervised machine learning, random forest, risk prediction

I. Introduction

Diabetes is a form of sickness that lasts for a long time and is chronic. People of all ages might be affected by this condition. It does not matter what age you are. This disease impacts our health status by preventing our bodies from converting the fuel we get from food into usable energy [1]. Diabetes can lead to a number of serious complications, including loss of vision, renal failure, heart attack, stroke, and amputation of lower limbs. In low- and middle-income nations, its rate of expansion has been picking up speed in recent years. Yet in countries with high incomes, growth rates have been rather stagnant. Diabetes is expected to affect around 642 million people in the world by the year 2040 [2].

Diabetes is a chronic metabolic condition marked by elevated blood sugar levels (glucose) [3]. According to the World Health Organization (WHO), over 422 million people worldwide suffer from diabetes, and this figure is projected to rise to 642 million by 2040. Diabetes must be diagnosed early in order to avoid or postpone the onset of major consequences such as blindness, kidney failure, and heart disease. Early-stage diabetes prediction is the process of identifying those at risk for developing diabetes prior to the onset of symptoms [4]. This is significant because early intervention and changes in lifestyle can prevent or postpone the onset of diabetes. Genetics, age, obesity, a sedentary lifestyle, and a poor diet are all risk factors for diabetes. By evaluating these risk factors, healthcare practitioners may identify patients who are at a greater risk of getting diabetes and offer them the proper therapies.

Many approaches, including blood glucose testing, glycated hemoglobin (A1C) tests, oral glucose tolerance tests, and random blood glucose tests, can be used to detect early-stage diabetes. In conjunction with other risk factors such as family history, body mass index (BMI), and blood pressure, these tests are frequently used to forecast an individual's probability of getting diabetes. In recent years, machine learning and artificial intelligence (AI) have also been used to construct prediction models capable of identifying persons at risk for getting diabetes based on their medical history, lifestyle, and genetics. Prediction of early-stage diabetes is a crucial aspect of diabetes care and prevention. By identifying patients at risk for acquiring diabetes and treating them with appropriate therapies, healthcare practitioners may lower the global burden of diabetes and enhance the quality of life for millions of persons [5].

The health sector is an immensely fascinating and rewarding area of research, as it directly impacts the well-being and quality of life of millions of people. In our particular field of study, we are focusing on diabetes patients and leveraging the power of real-time data to develop novel insights and solutions. One of the most exciting aspects of our research is the use of real-time data, which is a relatively uncommon approach in this field. By utilizing real-time data, we are able to gain a more accurate and comprehensive understanding of the factors that contribute to diabetes, as well as the most effective strategies for prevention and management. This is a highly motivating factor for us, as we are constantly pushing the boundaries of what is possible and discovering new and innovative ways to improve the lives of those living with diabetes. Ultimately, we are driven by the desire to make a meaningful impact in the lives of people

979-8-3503-0884-6/23/$31.00 ©2023 IEEE

affected by this condition, and we believe that our research has the potential to do just that.

This study progress by merging patient records from a number of different top-tier hospitals in Bangladesh. by using the dataset to train a machine learning algorithm. Since we based our model on information from lab reports of people with diabetes. As a result, the job will be flawless in every respect. Next, we construct a supervised learning model to evaluate our prediction performance. Now, the focus of our study is on evaluating the performance of the measurement models we've developed using various machine learning methods. Using the acquired data, we will thoroughly test the model's ability to reliably anticipate outcomes. Particularly, we will evaluate how effectively the algorithms respond to our dataset and compare the results to industry-standard benchmarks. The goal of this research is to develop a measuring model with extraordinary precision, which we believe will have a significant influence on diabetes research. By obtaining high levels of precision, we will be able to create more effective ways for predicting and controlling diabetes, therefore enhancing the lives of millions of people who are afflicted with this ailment. Our research has the potential to yield enormous advances, and we are working relentlessly to guarantee that our measurement models are as accurate and trustworthy as possible.

II. Literature Review

In [6], authors presented a new categorization scheme for diabetes evaluation and classification. Continuous glucose monitoring (CGM) is used to keep track of the patient's glucose level at predetermined intervals. The authors analyzed data on the Chinese population gleaned from clinical records housed at the People's Hospital of China. A total of 17 characteristics were retrieved using GSM, and then AdaBoost variant algorithms were utilized as a novel indicator for diabetes diagnosis and classification, with experimental findings of 90.3% accuracy. Metrics like average conversation length (ACC) and mean conversation content (MCC) were utilized to draw conclusions about the outcomes.

Gestational diabetes, a major risk factor for the development of type 2 diabetes, was identified by the authors in [7]. Increased body mass index (BMI) during pregnancy, certain racial/ethnic groups, and advanced age are all major risk factors. Diabetes mellitus (GDM) and other forms of diabetes during pregnancy were compared. ANOVA and SPSS were also used for statistical analysis. With WEKA, we built a prediction model utilizing methods like the naive Bayes classifier and J48. Pearson correlation coefficients were used to examine clinically relevant factors. Using a metabolomics signature, this study demonstrates the progression from gestational to type 2 diabetes.

The SVM machine learning method was proposed by the authors of [8] for the diagnosis of diabetes. In SVM, input is transformed using kernel functions applied to a huge multidimensional space. Both types of categorization methods employed 14 unique characteristics to differentiate between those with and without a diagnosis of diabetes, prediabetes, or any form of metabolic syndrome. ROC and other cross-validation methods were utilized to analyze performance. The best classification systems were bested by the RBF and linear kernel functions.

The authors of the research [9] advocated utilizing fuzzy SVM to detect diabetes at an early stage. The PID database was mined for a dataset consisting of eight characteristics. All eight qualities went through data pre-processing, however, only six were included in the final analysis. The greatest features were refined by filtering out the unnecessary ones with the help of feature selection and the F-score. The diabetes prediction categorization was then carried out using fuzzy SVM. For the purpose of analyzing HbA1 in diabetic patients using type 2 diabetes clinical records, a machine-learning logical regression model was presented.

Chun et al. [10] highlighted the increasing diabetes prevalence in Taiwan, with 2.18 million affected individuals. They underscored the urgency of prevention and holistic management due to the potentially fatal complications. Analyzing 15,000 women's data, their study employed machine learning models, identifying a two-class boosted decision tree as the most effective predictor of diabetes, achieving an AUC score of 0.991. Such insights are crucial for informed healthcare planning to mitigate the impending burden.

Tasin et al. [11] address diabetes as a global concern affecting millions. They develop a predictive model for diabetes using private female patient data from Bangladesh. Machine learning techniques are employed, including XGBoost with ADASYN for class imbalance. The proposed system achieves 81% accuracy, with explainable AI insights from LIME and SHAP frameworks. A web framework and Android app are also developed for instant predictions. This work highlights the potential of ML in early diabetes prediction.

Khaleel et al. [12] reviewed Diabetes as a metabolic disorder characterized by prolonged elevated blood sugar levels. Early prediction could mitigate severity. Machine learning's prominence in medicine inspired their model using various algorithms—Logistic Regression, Naïve Bayes, and K-nearest Neighbor—evaluated on precision, recall, and F1-measure. Applying PIDD dataset, their study achieved 94%, 79%, and 69% prediction rates. Notably, Logistic Regression outperformed other algorithms in diabetes prediction efficiency.

III. Proposed Methodology

Calculating nominal data from numerical data is a difficult task. In other words, text data is completely foreign to the machine learning algorithm. Numbers are all it knows, and that's about all. So, the textual information is not instantly useful, and it cannot have a measurable outcome. Our study, as far as we are aware, is predicated on precision. Given this, it's imperative that we have access to some sort of quantitative information. This means that we also have difficulties when trying to convert text data into numerical data. We've employed these pre-processed data in our algorithms after first implementing processing on nominal and transforming [13], the nominal data into the numeric data form. For this purpose, we have employed supervised machine learning methods. The process of operation is described in the next section. Machine Learning is the most effective way to predict dearly stage diseases [14]. The development of these Machine Learning algorithms is a major focus for many scientists. Nobody special, that's for sure. This means that we have experimented with a number of ML methods. As a result, we initially studied machine learning (ML) algorithms and python. Here we describe every machine learning algorithm that has ever been discovered.

A. Data Collection & Pre-Processing

When working with data that is missing, noisy, or otherwise inconsistent, data preparation is an essential step in the data mining process [15]. Data preparation encompasses a wide range of procedures, including as data cleansing, data defuzzification, data integration, processing data, data conversion, and so on, with the goal of consistently presenting data in a cohesive and proper form. The UCI repository [16] provides a variety of datasets, including diabetes data with 17 variables used for this case study. In this case, we use a dataset with 17 characteristics that indicate both patient and hospital outcomes. It consists of conventional therapeutic data. This dataset has been used to evaluate the efficacy of ensemble algorithms for predicting future outcomes. While working with data, some mining methods perform better when dealing with discrete qualities. Discrete characteristics are what define a category; they are also called nominal attributes. A category's ordinal features are those that define it and bear weight on its place in the hierarchy of categories. The term "discretization" refers to the transformation of a continuous variable into discrete categories. As the input values are actual, a discretize filter was used so that they may be organized into categories. In this work, we use 520 instances and 17 characteristics, one of which is a class attribute, to predict if an individual would get diabetes.

B. Data Description

We begin by making an effort to ascertain the nature of the information included inside our data structure, including the total number of rows and columns. There were in total 3837 records with 17 columns. We used data visualization to identify blanks. The number of True and False classes was then determined by visually representing each column separately. The proportion and total amount of data in each column have displayed separately. There are 3837 observations in the dataset, 3057 of which correspond to the true cases and 396 were false cases. Table 1 shows the description of the dataset.

TABLE I. DESCRIPTION OF DATASET

Attribute	Description
Age	Age of the individual
Gender	Gender of the individual (Male/Female)
Polyuria	Excessive urination
Polydipsia	Excessive thirst
Sudden weight loss	Rapid and unexplained weight loss
Weakness	General lack of strength or energy
Polyphagia	Excessive hunger or increased appetite
Genital thrush	Fungal infection in the genital area
Visual blurring	Blurred or unclear vision
Itching	Skin itching or irritation
Irritability	Easily becoming annoyed or agitated
Delayed healing	Slow healing of wounds or injuries
Partial paresis	Partial loss of muscle function or weakness
Muscle stiffness	Stiffness or inflexibility in muscles
Alopecia	Hair loss or baldness
Obesity	Excessive body weight
class	The class or category of the medical condition

C. Proposed Model Workflow

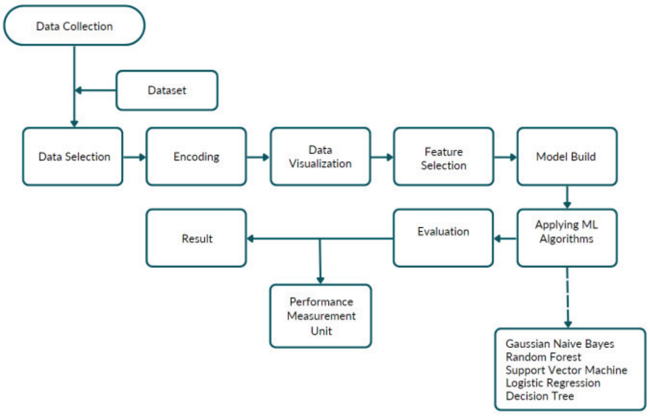


Fig. 1. Proposed model workflow

D. Machine Learning Model

Predicting individuals who will develop diabetes using machine learning is the most feasible option. From the research summarized in the Literature Review, it is obvious that machine learning and deep learning have been the primary methods of data analysis thus far. The field of machine learning, under which deep learning is typically included, is widely accepted as including this idea. In order to determine which machine learning technique was most effective on this novel dataset, five different approaches were tested. These techniques are classified under the headings of Random Forest [17], Decision Tree [18], Gradient Boosting [19], Logistic Regression [20], and GaussianNB [21]. Below, we'll have a brief survey of a few of these designs.

1) Decision Tree (DT)

For both categorization and prediction, nothing beats a Decision Tree. Each node in a decision tree represents an attribute test, each branch an outcome of that test, and each leaf node (the terminal node) a class label. DT is a powerful machine learning technique for predictive modeling and data categorization. Each hypothesis is represented by a node in the vector, and the endpoints of the vector give a predicted category or value. Because of this, the vector can progress in a forward direction. There are a few different ways DT might be structured. While DT does well when there are only a few classes and enough of data to train a model, it struggles when there are numerous classes and not enough data. DT works best in situations when there are few instances to study. As an added complication, training DTs may need a large amount of computer resources. [18].

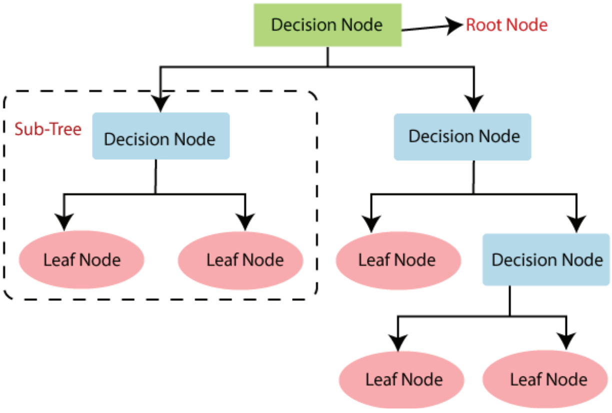


Fig. 2. Decision tree visualization

2) *Random Forest (RF)*

When it comes to machine learning, the supervised learning approach known as Random Forest has shown to be rather successful. Classification and regression are two ML applications that benefit from its utilization. The method relies on ensemble learning, which entails using several classifiers to increase model performance and tackle a more difficult issue. In Random Forest, there is a plethora of paths to take. The outcome is highly correlated with the density of the forest. The higher the sample size of trees, the more trustworthy the results. In RF, the classifier is often a C4.5 or J48. Bagging was one of the feature options for decision trees that Breiman presented in his article. RF is a form of classifier that requires human supervision. [17].

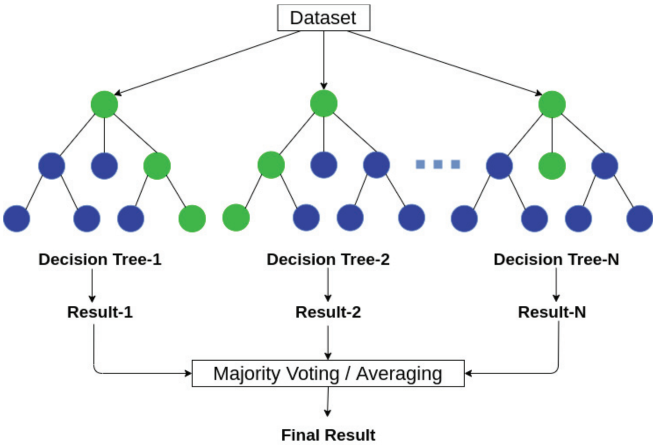


Fig. 3. Random Forest visualization

3) *Gradient Boosting (XGB)*

One well-liked boosting approach is called Gradient Boosting. To improve accuracy, gradient boosting uses successive predictors to compensate for the mistakes made by the one before it. In contrast to Adaboost, each predictor is trained with the residual errors of the predecessor as labels, rather than by adjusting the weights of the training cases. For example, CART serves as the foundation for the Gradient Boosted Trees method (Regression Trees). The XGBoost technique, which is an aggregation decision tree approach, is widely used in gradient boosting frameworks. While decision trees seem apparent at a glance, it may be more challenging to acquire a first-hand understanding of older tree-based algorithms [19].

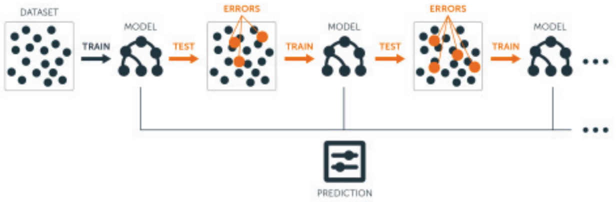


Fig. 4. Gradient Boosting visualization

4) *Logistic Regression (LR)*

In statistical analysis, logistic regression is used to make predictions about a binary result from a series of observations, such as yes or no. By examining the association between one or more independent variables, a logistic regression model can make predictions about a dependent data variable. A logistic regression might be used to foretell the success or failure of a political candidate in an election, or of a high school senior's application to a specific institution. These easy-to-understand binary options simplify choosing between two feasible solutions. Logistic regression is a type of supervised learning known as categorization. There are only discrete ways in which X can influence the attribute (or output) y at the center of the categorization problem. It's correct that logistic regression may be categorized among other types of regression models. A regression approach is built to estimate the probability that the input data is a 1. Logistic regression may be used to rapidly solve classification problems, such as those encountered in cancer diagnosis[20].

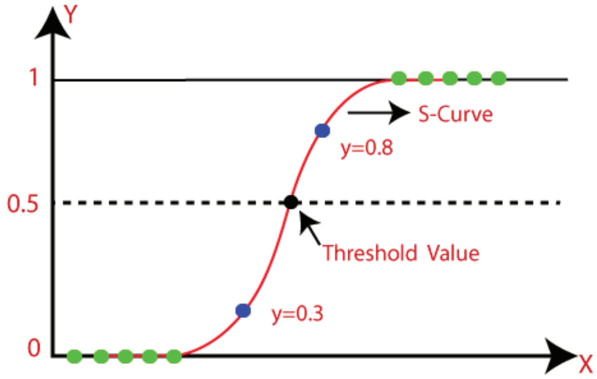


Fig. 5. Logistic Regression graph

5) *Gaussian NB*

The Gaussian naive Bayes algorithm is a technique for making probabilistic classifications based on the interrelationships among characteristics in the data set; it is derived from Bayes' theorem in statistics. This approach assumes that the value of a class variable is separate from any other attribute in the dataset, and proceeds with the operations when the algorithm is given such a variable. Each variable's property is assumed to have an equal and uncorrelated impact on the variable's output value by this classifier. Before calculating the variance and mean of the x (continuous attribute) value, the data are first classified into their respective classes [21].

IV. Experimental Result

While training or extrapolating findings, many individuals err. Increasing the complexity of the model can help reduce training errors since the training error rate reduces with increased complexity. Correct generalizations may be made less frequently with the help of the Bias-Variance Decomposition (Bias + Variance) technique. Training error reduction leading to test error increases is an example of overfitting. The F1-Score, recall, precision, and accuracy of a classification system are some measures by which it may be evaluated.

In order to determine how well their models performed, authors employed a broad variety of methods. Several studies utilized many indicators to evaluate performance, whereas others relied on just one. In this case, we measure the effectiveness of the job based on its accuracy, precision, recall, and F_1-Score. This four-factor framework works quite well for examining prediction data.

The capacity to appropriately recognize and categorize incidents is related to accuracy.

Equation 1 shows the formula of accuracy [15].

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

Specifically, accuracy in statistics is defined as the ratio of actual positive occurrences to the total predicted positive events. The mathematical expression of accuracy is given by Equation 2 [22].

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

How successfully the algorithm is able to identify persons who have cancer is quantified by a metric called "recall" [20]. Mathematically, recall is represented by Equation 3.

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

The term "harmonic mean" describes this method since it balances accuracy and memory. A version of the mathematical equation for the F_1 -Score [23] is given by Equation 4.

$$F_1 - \text{Score} = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}\right) \quad (4)$$

A. Result

It has been used with five different machine learning techniques. When contrasting the efficacy of different algorithms, a narrow gap is seen. When compared to the other five algorithms, we have seen that RF achieved the best accuracy which is 98% DT and XGB work similarly. They both achieved 96% accuracy. On the other hand, GNB didn't work well between these 5 algorithms. GNB achieved 88% model accuracy. In terms of F_1 −Score, DT performs well here. DT achieved 0.98 scores which is the best score. Table II shows the classification report of all algorithms.

TABLE II. CLASSIFICATION REPORT OF THIS MODEL

Algorithms	Precision	Recall	F_1 − Score	Accuracy
RF	0.979	0.980	0.979	0.983
DT	0.971	0.972	0.981	0.969
XGB	0.967	0.966	0.967	0.966
LR	0.918	0.916	0.917	0.917
GNB	0.883	0.882	0.882	0.883

In Fig. 6, AUC comparison of five machine learning algorithm has been shown.

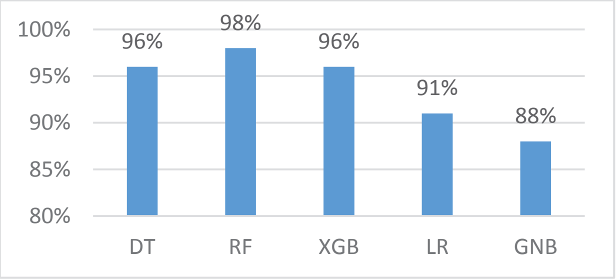
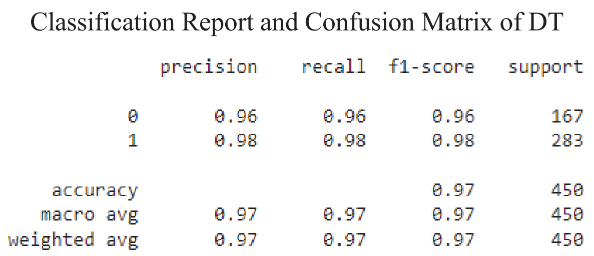


Fig. 6. AUC comparison of five machine learning algorithm.

Confusion matrices may be used to quickly and readily summarize the effectiveness of a classification system. Even though there are just two categories in the dataset, the classification may be erroneous if there is a large difference in the number of observations across categories. For further insight into the accuracy of the classification method, a confusion matrix can be computed (CM) [24].



Classification Report and Confusion Matrix of DT

```
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       167
           1       0.98      0.98      0.98       283

    accuracy                           0.97       450
   macro avg       0.97      0.97      0.97       450
weighted avg       0.97      0.97      0.97       450
```

Fig. 7. Classification report of DT

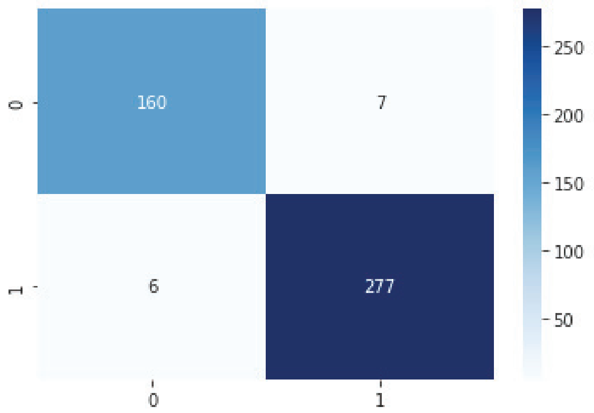
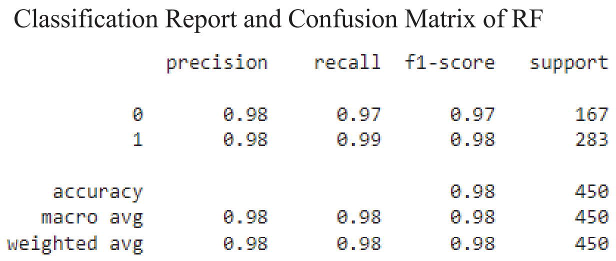


Fig. 8. Confusion Matrix of DT



Classification Report and Confusion Matrix of RF

```
              precision    recall  f1-score   support

           0       0.98      0.97      0.97       167
           1       0.98      0.99      0.98       283

    accuracy                           0.98       450
   macro avg       0.98      0.98      0.98       450
weighted avg       0.98      0.98      0.98       450
```

Fig. 9. Classification report of RF

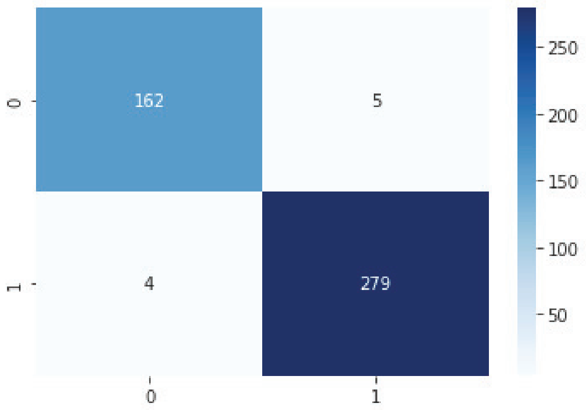
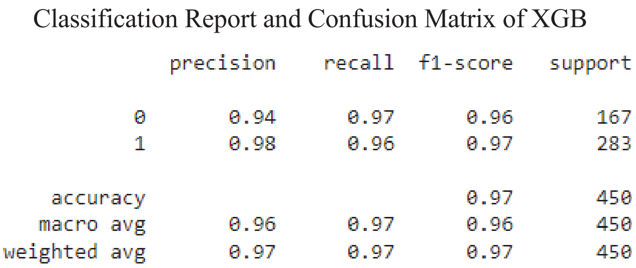


Fig. 10. Confusion Matrix of RF



Classification Report and Confusion Matrix of XGB

```
              precision    recall  f1-score   support

           0       0.94      0.97      0.96       167
           1       0.98      0.96      0.97       283

    accuracy                           0.97       450
   macro avg       0.96      0.97      0.96       450
weighted avg       0.97      0.97      0.97       450
```

Fig. 11. Classification report of XGB

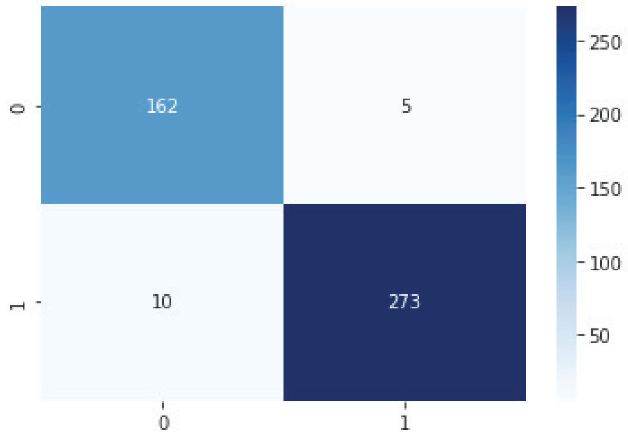
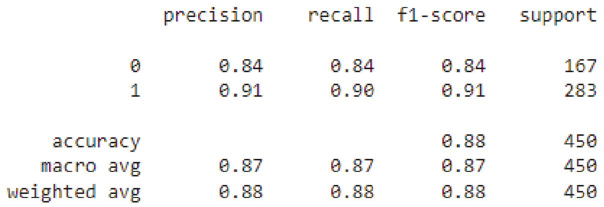


Fig. 12. Confusion Matrix of XGB

Classification Report and Confusion Matrix of GNB



```
              precision    recall  f1-score   support

           0       0.84      0.84      0.84       167
           1       0.91      0.90      0.91       283

    accuracy                           0.88       450
   macro avg       0.87      0.87      0.87       450
weighted avg       0.88      0.88      0.88       450
```

Fig. 13. Classification report of GNB

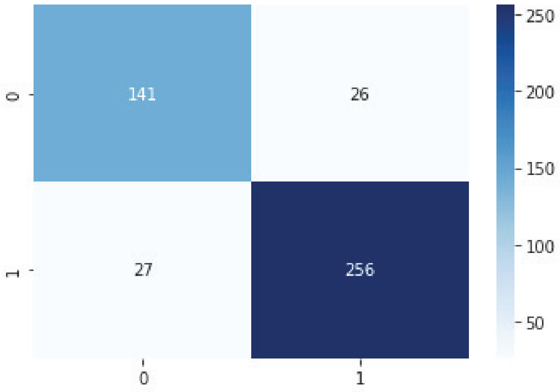


Fig. 14. Confusion Matrix of GNB

V. Discussion

The health effects of many prevalent disorders are severe. Diabetes is one of the most prevalent diseases nowadays. Many often say that diabetes is the cause of every other ailment. The elevated blood sugar is associated with the diabetes illness. All of our body's energy originates from the glucose in our blood. One of the most common complications of diabetes is blindness. Insulin, a hormone produced by the pancreas, is responsible for transporting glucose from the food we eat into the cells of the body, where it is utilized for energy. These days, these illnesses strike the majority of the global population. A significant number of patients are blind as a result of this reason. Insulin production might be faulty at times. As a result, our blood glucose levels rise dangerously quickly, posing serious health risks. There is currently no treatment for diabetes, therefore it is important to follow a set of guidelines to help keep you healthy. This research aims to evaluate the effectiveness of various machine learning algorithms for early-stage diabetes risk prediction in Bangladesh. The study uses real data from diabetic patients and employs five well-known machine learning algorithms to determine performance metrics. The Random Forest-based classifier was found to have the highest accuracy, at 98%, making it the superior algorithm for this application. The study highlights the potential of machine learning techniques in the medical field for early detection and prevention of diabetes, which can help reduce the prevalence of the disease and its associated health problems.

VI. Conclusions and Future Work

The use of machine learning algorithms for early-stage diabetes risk prediction shows great promise in improving healthcare outcomes in Bangladesh. The results of this research demonstrate that the Random Forest-based classifier is the most accurate method for this application, which can be a valuable tool for healthcare practitioners to identify patients at risk and provide early interventions. However, ethical considerations such as data privacy and bias must be carefully considered to ensure the responsible use of these technologies. Additionally, promoting healthy lifestyles and preventative measures should be an integral part of any sustainability plan for addressing the diabetes epidemic. Overall, this study highlights the potential of machine learning techniques to improve healthcare outcomes and contribute to broader public health goals. There is a vast array of algorithms for machine learning. To evaluate performance indicators, we use five well-known ML algorithms in this study: the Gaussian Naive Bayes classifier, the Random Forest classifier, the Support Vector Machine classifier, the Logistic Regression classifier, and the Decision Tree classifier. Utilizing actual data from Bangladeshi diabetes patients, these algorithms were constructed and assessed. We rely on the effectiveness of these techniques. Out of five distinct machine learning algorithms, the Random Forest-based approach had the best accuracy, at 98%.

For this research we have recommendations. Machine learning classification is being used in my work for developing the accuracy of the model. For a large number of datasets there are many algorithms and methods, datasets. So, that model will predict breast cancer detection. Recommendations are given below:

- More Accurate Breast Cancer Prediction Dataset.
- Try to create better classification models.
- Will try to implement better deep learning model.
- Try to find better execution of accuracy.

References

[1] J. Abdollahi and B. Nouri-Moghaddam, "Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction," *Iran Journal of Computer Science 2022 5:3*, vol. 5, no. 3, pp. 205–220, Mar. 2022, doi: 10.1007/S42044-022-00100-1.

[2] S. Islam Ayon and M. Milon Islam, "Information Engineering and Electronic Business," *Information Engineering and Electronic Business*, vol. 2, pp. 21–27, 2019, doi: 10.5815/ijieeb.2019.02.03.

[3] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Comput Sci*, vol. 165, pp. 292–299, Jan. 2019, doi: 10.1016/J.PROCS.2020.01.047.

[4] M. Saberi-Karimian *et al.*, "Data mining approaches for type 2 diabetes mellitus prediction using anthropometric measurements," *J Clin Lab Anal*, vol. 37, no. 1, p. e24798, Jan. 2023, doi: 10.1002/JCLA.24798.

[5] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, “Diabetes prediction using supervised machine learning,” *Procedia Comput Sci*, vol. 216, pp. 21–30, Jan. 2023, doi: 10.1016/J.PROCS.2022.12.107.

[6] G. Cappon, M. Vettoretti, G. Sparacino, and A. Facchinetti, “Continuous Glucose Monitoring Sensors for Diabetes Management: A Review of Technologies and Applications,” *Diabetes Metab J*, vol. 43, no. 4, pp. 383–397, Aug. 2019, doi: 10.4093/DMJ.2019.0121.

[7] J. Yang *et al.*, “Modifiable risk factors and long term risk of type 2 diabetes among individuals with a history of gestational diabetes mellitus: prospective cohort study,” *BMJ*, vol. 378, Sep. 2022, doi: 10.1136/BMJ-2022-070312.

[8] R. Srivastava and R. K. Dwivedi, “A Survey on Diabetes Mellitus Prediction Using Machine Learning Algorithms,” *Lecture Notes in Networks and Systems*, vol. 321, pp. 473–480, 2022, doi: 10.1007/978-981-16-5987-4_48/COVER.

[9] M. R. Rajput and S. S. Khedgikar, “Diabetes prediction and analysis using medical attributes: A Machine learning approach”, doi: 10.37896/JXAT14.01/314405.

[10] C. Y. Chou, D. Y. Hsu, and C. H. Chou, “Predicting the Onset of Diabetes with Machine Learning Methods,” *Journal of Personalized Medicine 2023, Vol. 13, Page 406*, vol. 13, no. 3, p. 406, Feb. 2023, doi: 10.3390/JPM13030406.

[11] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, “Diabetes prediction using machine learning and explainable AI techniques,” *Healthc Technol Lett*, vol. 10, no. 1–2, pp. 1–10, Feb. 2023, doi: 10.1049/HTL2.12039.

[12] F. Alaa Khaleel and A. M. Al-Bakry, “Diagnosis of diabetes using machine learning algorithms,” *Mater Today Proc*, vol. 80, pp. 3200–3203, Jan. 2023, doi: 10.1016/J.MATPR.2021.07.196.

[13] M. A. Sheakh, M. Sazia Tahosin, M. M. Hasan, T. Islam, O. Islam, and M. M. Rana, “Child and Maternal Mortality Risk Factor Analysis Using Machine Learning Approaches,” *ISDFS 2023 - 11th International Symposium on Digital Forensics and Security*, 2023, doi: 10.1109/ISDFS58141.2023.10131826.

[14] T. Islam, A. Vuyia, M. Hasan, and M. M. Rana, “Cardiovascular Disease Prediction Using Machine Learning Approaches,” *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pp. 813–819, Apr. 2023, doi: 10.1109/CISES58720.2023.10183490.

[15] J. A. Swets, “Measuring the Accuracy of Diagnostic Systems,” *Science (1979)*, vol. 240, no. 4857, pp. 1285–1293, 1988, doi: 10.1126/SCIENCE.3287615.

[16] “Early stage diabetes risk prediction dataset. - UCI Machine Learning Repository.” https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset (accessed Aug. 22, 2023).

[17] S. J. Rigatti, “Random Forest,” *J Insur Med*, vol. 47, no. 1, pp. 31–39, Jan. 2017, doi: 10.17849/INSM-47-01-31-39.1.

[18] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, “An introduction to decision tree modeling,” *J Chemom*, vol. 18, no. 6, pp. 275–285, Jun. 2004, doi: 10.1002/CEM.873.

[19] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artif Intell Rev*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/S10462-020-09896-5/METRICS.

[20] T. Islam, A. Kundu, N. Islam Khan, C. Chandra Bonik, F. Akter, and M. Jihadul Islam, “Machine Learning Approaches to Predict Breast Cancer: Bangladesh Perspective,” *Smart Innovation, Systems and Technologies*, vol. 302, pp. 291–305, 2022, doi: 10.1007/978-981-19-2541-2_23/COVER.

[21] R. Pushpakumar, R. Prabu, M. Priscilla, P. S. Renisha, R. T. Prabu, and U. Muthuraman, “A Novel Approach to Identify Dynamic Deficiency in Cell using Gaussian NB Classifier,” *7th International Conference on Communication and Electronics Systems, ICCES 2022 - Proceedings*, pp. 31–37, Feb. 2022, doi: 10.1109/ICCES54183.2022.9835813.

[22] T. Islam, M. A. Hosen, A. Mony, M. T. Hasan, I. Jahan, and A. Kundu, “A Proposed Bi-LSTM Method to Fake News Detection,” *2022 International Conference for Advancement in Technology, ICONAT 2022*, 2022, doi: 10.1109/ICONAT53423.2022.9725937.

[23] Md. T. Islam, T. Ahmed, A. B. M. Raihanur Rashid, T. Islam, Md. S. Rahman, and Md. Tarek Habib, “Convolutional Neural Network Based Partial Face Detection,” in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, IEEE, Apr. 2022, pp. 1–6. doi: 10.1109/I2CT54291.2022.9825259.

[24] Md. S. H. Talukder, R. Bin Sulaiman, M. R. Chowdhury, M. S. Nipun, and T. Islam, “PotatoPestNet: A CTInceptionV3-RS-Based Neural Network for Accurate Identification of Potato Pests,” *Smart Agricultural Technology*, p. 100297, Jul. 2023, doi: 10.1016/J.ATECH.2023.100297.