

Automatic diabetes prediction with explainable machine learning techniques

Adiba Haque, Sanjida Islam, Nusrat Rahim Mim, Sabrina Mannan Meem, Ananya Saha, Riasat Khan

Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

Article Info

Article history:

Received Mar 22, 2024

Revised Jul 7, 2024

Accepted Jul 26, 2024

Keywords:

Artificial intelligence

Diabetes prediction

Explainable artificial intelligence

Machine learning

Metabolic disorder

Synthetic oversampling technique

ABSTRACT

Diabetes is a metabolic disorder caused by various genetic, physiological and behavioral factors. It occurs due to an imbalance in the body's insulin processing, which results in elevated blood sugar levels. Its early diagnosis can alleviate the risk of other deadly diseases. The onset and accurate detection of diabetes can decrease the progression of different complications and dysfunction of tissues. The principal objective of this article is to utilize machine learning approaches to predict the existence of diabetes in female patients at a primary stage. Multiple machine learning, including ensemble classifiers with the Pima Indian dataset and a private dataset obtained from a local Bangladeshi hospital, are used in this work. We employed feature scaling, synthetic oversampling technique (SMOTE), and hyperparameter optimization with GridSearchCV to get the best performance from different machine learning algorithms. The support vector machine (SVM) with the SMOTE framework and default hyperparameters achieved the accuracy and F1 score of 87% and 91%, respectively. The accuracy and F1 score of the SVM model improved to 95% and 91%, respectively, with hyperparameter optimization. Finally, explainable artificial intelligence with the local interpretable model-agnostic explanations (LIME) is employed to illustrate the predictability of the SVM technique.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Riasat Khan

Department of Electrical and Computer Engineering, North South University

Plot: 15, Block: B, Bashundhara, Baridhara, Dhaka-1229, Bangladesh

Email: riasat.khan@northsouth.edu

1. INTRODUCTION

When the body's insulin is used inadequately, and consequently, the pancreas fails to generate adequate insulin, then an inmedicable disease occurs known as diabetes [1]. There are different types of diabetes characterized by hyperglycemia, for instance, type one, type two, and gestational diabetes. Prediabetes is also considered another type of diabetes. Sometimes people have a glucose level that is more excess than standard but not that much excess to type 2 diabetes. This condition is called prediabetes [2]. Worldwide 537 million individuals aged from 20 to 79 years have been affected by diabetes, according to a report by World Health Organization (WHO) published in 2021. The information anticipated that by 2030 and 2045, the rate would be increased to 643 million and 783 million, respectively [2]. In 2019, approximately 8.40 million adults had diabetes in Bangladesh, which is anticipated to expand to almost 15 million by 2045. 3.80 million people were expected to have prediabetes in 2019. 8.2% of rural women and 12.9% of females in urban areas of Bangladesh are affected by gestational diabetes mellitus [3]. The treatment of diabetes varies in steps, with what amount of insulin the body makes and how properly the body can use available insulin. Diabetes is not curable, yet it is controllable. Diabetes care is associated with adopting a healthy lifestyle, restricted diet, weight control and regular physical activity.

Accurate and prompt prediction of diabetes is a concern. Notable works have been done on the automatic identification of diabetes utilizing various machine learning approaches. The automated prediction of these works is expected to come up with a helpful referencing tool and preliminary judgment for clinicians to predict diabetes early on and reduce the workload of healthcare professionals. Some significant works based on diabetes prediction have been described briefly in the following paragraphs.

Many of these studies used different open-source datasets, particularly the Pima Indian dataset. For instance, Abdulhadi and Al-Mousa [4] conducted machine learning approaches to determine type 2 diabetes in females at the primary stage. The authors used the Pima Indian dataset containing 768 instances with nine attributes, but there were many missing values from several cases. Hence the null values were restored with the mean value, and the dataset was standardized using a standard scalar. The authors tested different classifiers and random forest achieved the highest accuracy of 82%, making it the most efficient model. Hasan *et al.* [5] used various feature selection approaches to create a tree-based prediction model for the early detection of diabetes in female patients. The authors used several steps for data preprocessing, i.e., handling missing values, feature selection, oversampling, and feature scaling. The authors used decision tree, random forest, extra trees, and adaptive boosting (AB) frameworks. The extra trees model provided the highest accuracy and F1 score of 88.3 percent and 0.877, respectively. Khanam and Foo [6] applied different machine learning approaches and artificial neural network (ANN) on the Pima Indian dataset to predict diabetes. The author applied the traditional data preprocessing methods using the WEKA tool and the K-fold cross-validation technique. ANN obtained the maximum accuracy of 86% among all the proposed models. Chang *et al.* [7] observed the random forest model to obtain the highest accuracy of 79.57% in predicting whether the patient is diabetic or non-diabetic. Bano and his team [8] applied support vector machine (SVM), ANN, decision tree, logistic regression, and farthest first clustering machine learning approaches on the Pima Indian dataset to predict diabetes with better accuracy. The authors preprocessed data by applying efficient feature selection modalities and also split the dataset into train and test data. Finally, after applying machine learning approaches to the preprocessed dataset, they got the best accuracy of almost 90% from the farthest first approach. Naz and Ahuja [9] used the Pima Indian dataset and deployed several machine learning methods. For better prediction, the authors used synthetic oversampling and rapid mining studio for preprocessing and rudimentary prediction. The obtained accuracies for different models ranged from 0.90 to 0.98. ANN-based deep learning technique achieved the best result with 0.98 accuracy. Kumari *et al.* [10] applied an ensemble learning technique with a soft vote classifier to boost the prediction performance. The authors preprocessed various features in the public dataset using encoding labels and min-max normalization approaches. The soft voting ensemble technique produced the best performances: an F1 score of 0.806, a precision of 0.7348, 0.7145 recall, and 0.9702 classification accuracy. Butt *et al.* [11] used various machine learning and tree-based ensemble algorithms to classify diabetes. Multilayer perceptron (MLP) was fine-tuned because of outperforming other algorithms with an accuracy of 86.08%.

Some of the articles employed custom datasets collected from various sources. Islam *et al.* [12] introduced two latest techniques to derive important features from the oral glucose tolerance test. Identifying the best segments and most critical risk factors that account for the further advancement of diabetes is a crucial contribution of the authors. The authors used data from san antonio heart study (SAHS). The arithmetical mean of the appropriate variable was used to fill in missing values in the raw dataset. The employed machine learning models were trained and tested using a 10-fold CV approach. The probability of a person having type 2 diabetes in the upcoming 7-8 years is predicted by the naïve Bayes approach with an accuracy of 95.94%. Pustozarov *et al.* [13] used their unique dataset collected from a Russian medical research center. The dataset includes 3,240 meal recordings and their related postprandial glycemic responses (PPGR) from patients. The authors employed gradient boosting models for prediction, trained with hyperparameter tuning and cross-validation. The most important factors influencing the PPGR are found to be glycemic load, carbohydrate count, and meal style. When the model did not use data on the current glucose level, the value for a person's correlation was 0.631, and the mean absolute error was 0.373 mmolL⁻¹. When the model used data on the current glucose level, the value for a person's correlation was 0.644, and the mean fundamental error was 0.371 mmolL⁻¹. When the model utilized data on continual blood glucose trends before the meal, the value for a person's correlation was 0.704, and the mean absolute error was 0.341 mmolL⁻¹. Azbeg *et al.* [14] employed Pima Indian dataset and a unique diabetes database from the Frankfurt Germany Hospital to develop a novel strategy for effectively predicting diabetes. The proposed model's accuracy was 99.5% for the Frankfurt dataset, 99.5% for the Pima Indian dataset, and 99.8% for the combined dataset, all based on an adaptive random forest method. For secure data collection and archival, the authors combined interplanetary file system (IPFS) distributed framework and blockchain with IoT medical devices, strengthening and improving the mechanism's consistency.

From the above paragraphs, we can decipher that extensive research has been done on automatic diabetes prediction. Various machine learning approaches were applied in these works to predict diabetes accurately. The majority of these studies used different open-source datasets without using any comprehensible

artificial intelligence techniques to analyze the predictions made by machine learning models. As a result, a combination of open-source and custom datasets and explainable artificial intelligence approaches have been proposed in this work.

This article used machine learning techniques to predict the possible presence of type 2 diabetes at an early stage. The significant contribution of this work can be listed as:

- A custom dataset of female diabetes patients collected from a local medical center in Bangladesh has been introduced to the scientific community. This dataset has been combined with the public Pima Indian dataset.
- Mean imputation, interquartile range (IQR)-based outlier detection and synthetic oversampling technique (SMOTE) techniques have been used in the data preprocessing stage.
- GridSearchCV hyperparameter optimization method has been applied for various machine learning and ensemble approaches.
- Explainable artificial intelligence library local interpretable model-agnostic explanations (LIME) is utilized to interpret the results and determine the main contributing factors that impact the prediction process, making this research more reliable.

The novelty of this work is the application of explainable machine learning techniques on a unique local dataset of female patients of Bangladesh.

2. METHOD

This section illustrates the dataset used, preprocessing techniques, applied machine learning models, and overall working sequences of the proposed automatic diabetes identification system. The workflow of the proposed diabetes prediction system starts with the merging of the Pima Indian dataset and a private dataset from Choto Gobra Community Clinic of Tangail, Bangladesh, followed by dataset preprocessing and data split into training and test sets. The training data undergoes SMOTE for balancing and is fed into various machine learning models with hyperparameter tuning approaches.

2.1. Dataset

The machine learning model has been trained and tested using a merged dataset which included the Pima Indian dataset [15] and a custom local dataset. The private dataset has been collected from Choto Gobra Community Clinic of Tangail, Bangladesh. The dataset includes 95 instances of female patients. It comprises various attributes, i.e., weight, height, blood pressure, glucose, age and number of pregnancies. The Pima Indian dataset contains 863 cases, separated into two classes with eight distinct attributes. To maintain consistency between the two datasets, we have used the shared five features, pregnancy, glucose, blood pressure, body mass index (BMI) and age. Table 1 illustrates the different features and outcomes of the merged dataset used in this work. As the table indicates, the combined dataset has five features and has an output that shows whether the person has diabetes (Yes) or not (No).

Table 1. Characteristics of the merged dataset used in this work

Feature	Description
Number of records (public)	768
Number of records (private)	95
Total Number of records (merged)	863
Total number of attributes	5
Attributes 1: Pregnancies	Range: 0 to 17
Attributes 2: Glucose (mg/dL)	Range: 0 to 199
Attributes 3: Blood pressure (mm Hg)	Range: 0 to 122
Attributes 4: BMI	Ranges: 0 to 67.10
Attributes 5: Age (years)	Range: 21 to 81
Outcome	Categorical: Yes (Diabetes), No (Healthy)

2.2. Dataset preprocessing

Medical data in the actual world is incoherent, bizarre, and complex. It may have null values, incomplete data, and outliers. Data preprocessing is one element that affects any classification system's performance. Classification inaccuracy will be elevated if the data quality is not facilitated [16].

In this work, we have attempted to handle the data as efficiently as possible. As a result, we went through various efficient preprocessing processes. In the merged dataset, some values were missing for several instances. For instance, it makes no sense to have zero blood pressure. The dataset has a relatively small number of cases. As a result, we did not exclude any instances with null values and instead used the mean imputation

method. Again, the maximum values in the sample appeared to be too high, e.g., a maximum BMI of 67.10 can be considered an outlier. We used IQR techniques to solve this problem of outlier detection. We also applied feature scaling in this work. The process of normalizing the independent feature values within a given range is known as feature scaling. Feature scaling is one of the most important data preprocessing steps [17]. The dataset had to be standardized since it had varied scales. The values of correlation between different attributes and final outcome (class) of the combined dataset, which range from 0 to 1, are shown in Table 2.

Table 2. Values of correlation of various features and outcome of the merged dataset

	Pregnancies	Glucose	BP	BMI	Age	Output
Pregnancies	1.00	0.129	0.141	0.018	0.544	0.222
Glucose	0.129	1.00	0.153	0.221	0.263	0.467
BP	0.141	0.153	1.00	0.282	0.239	0.065
BMI	0.017	0.221	0.282	1.00	0.036	0.292
Age	0.544	0.264	0.239	0.036	1.00	0.238
Output	0.221	0.467	0.065	0.292	0.239	1.00

2.3. Preparing machine learning models

We have used various machine learning methods in this research. Before applying different models, we employed a few ways to get the best results out of each model. We applied SMOTE, hyperparameter optimization with GridSearchCV, and explainable artificial intelligence approaches to get the best accuracy out of the machine learning algorithms.

- SMOTE: The SMOTE creates new synthetic observations for the minority category samples from the nearest neighbors in its corresponding feature space [18]. The primary purpose of this approach is to solve problems that occur when using an imbalanced dataset. It is an improved version of the traditional oversampling technique. The appropriate way of applying SMOTE directly on the training data set instead of the validation set. Applying SMOTE directly on the entire dataset creates new samples that would also appear in the validation or testing data set, giving us misleading results.
- GridSearchCV: For any machine learning model, we aim for the highest level of accuracy and optimal hyperparameters. In this work, we have used hyperparameter tuning with GridSearchCV. A hyperparameter from GridSearchCV is used to fine-tune the model in the specified range of all possible combinations of various hyperparameters [19]. When there are many parameters to tune, and the dataset is more extensive than its created computational issues and RandomizedSearchCV library is used. Since we have a small dataset, applying GridSearchCV helped to obtain accurate and robust results.
- Explainable artificial intelligence: The three key components of explainable artificial intelligence are forecasting accuracy, decision understanding, and traceability. Model-agnostic interpretability refers to the ability of LIME to explain why a machine learning model produces a particular result (outcome) for a given input [20]. The LIME interpretable artificial intelligence method helps illuminate a machine learning model and make predictions understandably.

2.4. Applied machine learning models

The following paragraphs depict brief descriptions of the employed models in this work.

- SVM: SVM belongs to the first of these three categories, i.e., supervised learning [21]. Classifying objects is one of the most basic machine learning problems, and SVM is one of the best classification methods. This model outperforms other techniques like neural networks in terms of processing speed and performance. It transforms data using kernel tricks. It calculates the ideal boundary between probable output through this process.
- Random forest: random forest is a collection of models that work together as an ensemble. Random forest is a multifunctional machine-learning method. It has been used in this article to predict diabetes and its effectiveness [22]. Random forest, unlike the decision tree method, generates a large number of decision trees. Every tree in the random forest gives categorization output and ‘vote’ when the random forest is expecting a new object based on some characteristics. The forest’s final output will be the most significant number in taxonomy.
- Decision tree: decision tree is the most effective and extensively used categorization and prediction tool [23]. One of the most common techniques for regression and classification is the decision tree. In this machine learning model, a class label is stored by each leaf node, and a test result is represented by each branch. The decision tree model’s tree structure can be utilized to explain the process of categorizing instances based on the impurity calculation of each feature.
- K-nearest neighbor (KNN): KNN is one of the most straightforward classifiers for solving classification problems [24]. This is a lazy learner algorithm and a non-parametric method.

- XGBoost classifier: XGBoost is an upgraded version of gradient boosting that stands for maximum gradient boosting [25]. This algorithm’s primary goal is to improve competition consistency and model performance. It has several features.
- AdaBoost classifier: The most widely used boosting technique for binary classification is called AdaBoost. It is a sequential learning process that means one tree is a previously dependent tree. If multiple models are implemented sequentially as M1, M2, and M3, it has a process of assembling, then M2 will depend on M1; similarly, M3 will depend on M2. Here all the models are dependent on each other. In AdaBoost, trees are not fully grown; they consist of one root and two leaves, referred to as stumps. It is advantageous to combine several weak classifiers into one strong classifier.

A merged dataset has been used in this work combining the Pima Indian and our collected datasets, illustrated in Figure 1. Necessary preprocessing techniques have been performed in the merged dataset, e.g., mean imputation for missing entries, feature scaling with min-max normalizer, and IQR-based outlier detection. Next, a stratified holdout validation approach with a 9:1 ratio was used to divide the dataset into training and test samples. We applied synthetic oversampling and GridSearchCV hyperparameter optimization approaches on train data. After that, we applied different machine learning techniques to create automatic prediction models. We evaluated each model’s predictive performance using different evaluation parameters. The best-performing models’ prediction performance has been represented using explainable artificial intelligence.

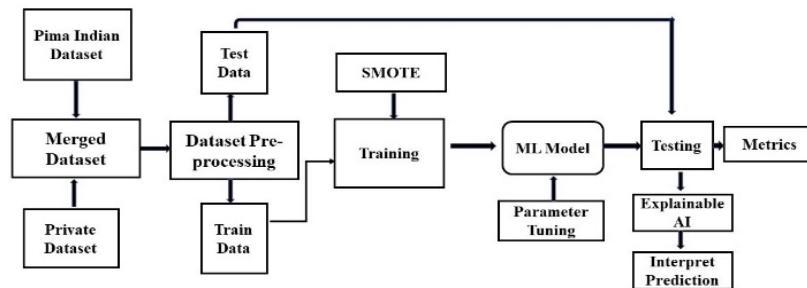


Figure 1. Working process of the proposed diabetes prediction system

3. RESULTS AND DISCUSSION

In this research, we used seven machine learning approaches in the combined dataset of 863 instances (Pima Indian and custom datasets) and five features. Table 3 depicts the performance metrics of various classifiers for default parameters and SMOTE technique. According to this table, SVM outperforms all the machine learning models with 87% accuracy and 91% F1 score.

Table 3. Performance metrics of various classifiers for default parameters and SMOTE

Classifier	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
SVM	91	91	91	87
Random forest	73	60	66	78
Logistic regression	66	78	72	74
Decision tree	56	56	56	70
KNN	69	45	55	73
XGBoost	66	62	64	79
AdaBoost	69	42	52	74

Table 4 demonstrates the performance metrics of various classifiers with GridSearchCV hyperparameter optimization and SMOTE. It states that both SVM and random forest achieve the highest precision and F1 score of 91%. Overall, the SVM model demonstrated the best performance with 95% accuracy and 91% F1 coefficient. The accuracy and F1 score improved with an average of 12% and 10%, respectively, after using the optimized hyperparameters. Figure 2 shows the accuracy of the machine learning models in the form of a bar graph with default parameters and SMOTE technique. It indicates that SVM has the best accuracy of 87% for the merged dataset.

The accuracy of the machine learning models is displayed in Figure 3 as a bar chart using GridSearchCV and SMOTE methods. According to this figure, the accuracy improved for all machine learning models. Figure 4 presents an illustration of the prediction interpretation of the SVM model using the LIME

explainable artificial intelligence framework. Since the SVM model with optimized hyperparameters and the SMOTE approach performed the best, it was utilized to evaluate the LIME prediction findings. According to this figure, the SVM model predicted diabetes (label: 1) for the individual patient with 96% confidence. The diabetes case is anticipated because of the age of less than 24, glucose level greater than 100 mg/dL and number of pregnancies greater than 1. Diabetes cases are anticipated because the age of the person is under 24, the blood glucose level is greater than 100 mg/dL, and there has been more than one pregnancy. Table 5 compares the proposed automatic diabetes prediction system with existing works. This table shows that our proposed model has been proven to be highly accurate compared to other works found in the literature.

Table 4. Performance metrics of various classifiers with optimized hyperparameters and SMOTE

Classifier	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
SVM	91	91	91	95
Random forest	91	91	91	92
Logistic regression	69	100	81	87
Decision tree	77	85	81	86
KNN	86	55	67	85
XGBoost	74	62	67	82
AdaBoost	59	68	63	76

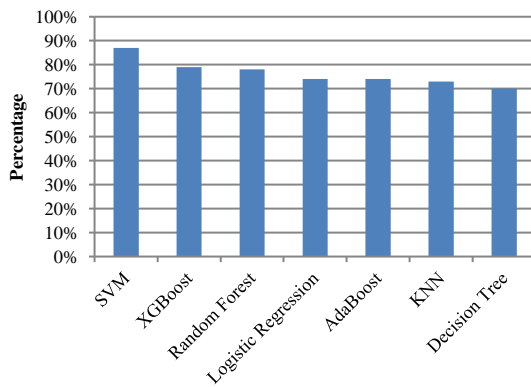


Figure 2. Validation accuracy of various employed machine learning models with default hyperparameters and SMOTE technique

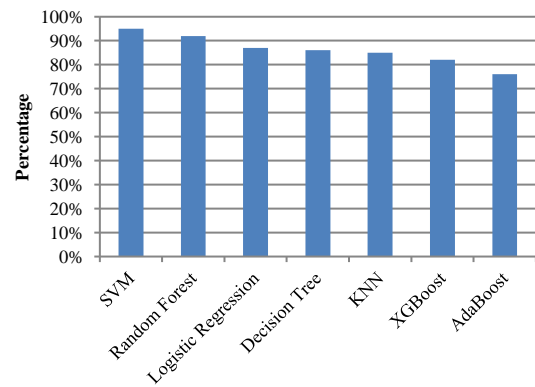


Figure 3. Validation accuracy of various employed machine learning models with hyperparameter tuning (GridSearchCV) and SMOTE technique

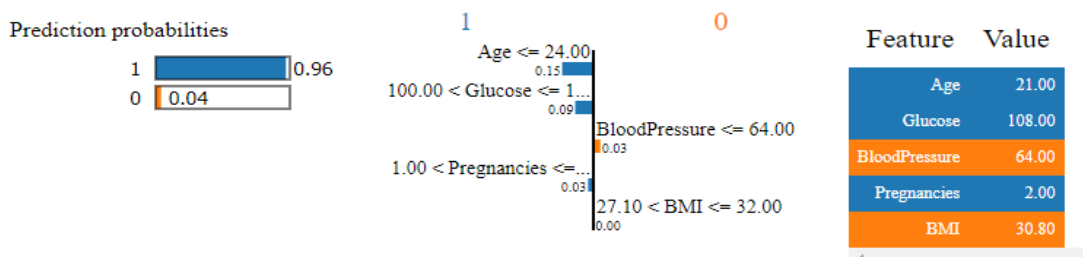


Figure 4. SVM model’s prediction interpretation with LIME explainable artificial intelligence

Table 5. Comparison of the proposed system with other works

Reference	Dataset	Accuracy and best-performed model	F1 score (%)
[5]	Pima Indian	88.3%: Extra trees	87.73
[6]	Pima Indian	86%: ANN	78.8
[7]	Pima Indian	79.57%: Random forest	85.17
[9]	Pima Indian	98.07%: ANN	94.72
[10]	Pima Indian	79.08%: Soft voting classifier	71.56
[11]	Pima Indian	87.26%: LSTM	N/A
This work	Pima Indian + Private dataset	95%: SVM	91





4. CONCLUSION

Diabetes mellitus is a severe metabolic disorder with increased cases worldwide due to the development of modern lifestyles. This work aims to build a model using various supervised machine learning methodologies to assist doctors in the early detection of diabetes, enhancing patient quality of life. The IQR approach of outlier detection, min-max normalizer-based feature scaling, and mean imputation for missing data have been used. The experimental outcomes illustrate that SVM performed better than the other approaches with the SMOTE technique and optimized hyperparameters. The prediction provided by the machine learning models has been interpreted by the LIME explainable artificial intelligence framework. In the future, an extensive dataset with diverse cohorts of patients and comprehensive features can be used.





REFERENCES

- [1] A. T. Kharroubi and H. M. Darwish, "Diabetes mellitus: The epidemic of the century," *World Journal of Diabetes*, vol. 6, no. 6, 2015, doi: 10.4239/wjd.v6.i6.850.
- [2] S. Schlesinger *et al.*, "Prediabetes and risk of mortality, diabetes-related complications and comorbidities: umbrella review of meta-analyses of prospective studies," *Diabetologia*, vol. 65, no. 2, pp. 275–285, Feb. 2022, doi: 10.1007/s00125-021-05592-3.
- [3] "Community-based detection and surveillance of gestational diabetes," *World Diabetes Foundation*, 2020. [Online]. Available: <https://www.worlddiabetesfoundation.org/projects/bangladesh-wdf15-962>
- [4] N. Abdulhadi and A. Al-Mousa, "Diabetes detection using machine learning classification methods," in *2021 International Conference on Information Technology (ICIT)*, Jul. 2021, pp. 350–354, doi: 10.1109/ICIT52682.2021.9491788.
- [5] S. M. M. Hasan, M. F. Rabbi, A. I. Champa, and M. A. Zaman, "An effective diabetes prediction system using machine learning techniques," in *2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)*, Nov. 2020, pp. 23–28, doi: 10.1109/ICAICT51780.2020.9333497.
- [6] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/j.ict.2021.02.004.
- [7] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Computing and Applications*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.
- [8] B. Farhana, K. Munidhanalakshmi, and D. R. M. Mohana, "Predict diabetes mellitus using machine learning algorithms," *Journal of Physics: Conference Series*, vol. 2089, no. 1, Nov. 2021, doi: 10.1088/1742-6596/2089/1/012002.
- [9] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *Journal of Diabetes and Metabolic Disorders*, vol. 19, no. 1, pp. 391–403, 2020, doi: 10.1007/s40200-020-00520-5.
- [10] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.
- [11] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine learning based diabetes classification and prediction for healthcare applications," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–17, Sep. 2021, doi: 10.1155/2021/9930985.
- [12] M. S. Islam, M. K. Qaraqe, S. B. Belhaouari, and M. A. Abdul-Ghani, "Advanced techniques for predicting the future progression of type 2 diabetes," *IEEE Access*, vol. 8, pp. 120537–120547, 2020, doi: 10.1109/ACCESS.2020.3005540.
- [13] E. A. Pustozarov *et al.*, "Machine learning approach for postprandial blood glucose prediction in gestational diabetes mellitus," *IEEE Access*, vol. 8, pp. 219308–219321, 2020, doi: 10.1109/ACCESS.2020.3042483.
- [14] K. Azbeg, M. Boudhane, O. Ouchetto, and S. J. Andaloussi, "Diabetes emergency cases identification based on a statistical predictive model," *Journal of Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00582-7.
- [15] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, no. 1–2, pp. 1–10, Feb. 2023, doi: 10.1049/htl2.12039.
- [16] M. N. I. Suvon, S. C. Siam, M. Ferdous, M. Alam, and R. Khan, "Masters and doctor of philosophy admission prediction of bangladeshi students into different classes of universities," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 4, pp. 1545–1553, Dec. 2022, doi: 10.11591/ijai.v11.i4.pp1545-1553.
- [17] P. Rajendra and S. Latifi, "Prediction of diabetes using logistic regression and ensemble techniques," *Computer Methods and Programs in Biomedicine Update*, vol. 1, 2021, doi: 10.1016/j.cmpbup.2021.100032.
- [18] A. Desiani, S. Yahdin, A. Kartikasari, and I. Irmeilyana, "Handling the imbalanced data with missing value elimination SMOTE in the classification of the relevance education background with graduates employment," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 2, pp. 346–354, Jun. 2021, doi: 10.11591/ijai.v10.i2.pp346-354.
- [19] R. Haque, S. Ho, I. Chai, and A. Abdullah, "Parameter and hyperparameter optimisation of deep neural network model for personalised predictions of asthma," *Journal of Advances in Information Technology*, vol. 13, no. 5, pp. 512–517, Oct. 2022, doi: 10.12720/jait.13.5.512-517.
- [20] R. Siddiqua, N. Islam, J. F. Bolaka, R. Khan, and S. Momen, "AIDA: Artificial intelligence based depression assessment applied to Bangladeshi students," *Array*, vol. 18, p. 100291, 2023, doi: 10.1016/j.array.2023.100291.
- [21] A. Y. A. Amer, "Global-local least-squares support vector machine (GLocal-LS-SVM)," *PLoS ONE*, vol. 18, no. 4, pp. 1–18, 2023, doi: 10.1371/journal.pone.0285131.
- [22] E. Asamoaha, G. B. M. Heuvelinka, I. Chairie, P. S. Bindrabanf, and V. Logah, "Random forest machine learning for maize yield and agronomic efficiency prediction in Ghana," *Heliyon*, vol. 10, no. 17, 2024, doi: 10.1016/j.heliyon.2024.e37065.
- [23] I. D. Mienye and N. Jere, "A survey of decision trees: concepts, algorithms, and applications," *IEEE Access*, vol. 12, pp. 86716–86727, 2024, doi: 10.1109/ACCESS.2024.3416838.
- [24] T. A. Assegie, "An optimized K-nearest neighbor based breast cancer detection," *Journal of Robotics and Control (JRC)*, vol. 2, no. 3, pp. 115–118, May 2020, doi: 10.18196/jrc.2363.
- [25] N. J. Riya, M. Chakraborty, and R. Khan, "Artificial intelligence-based early detection of dengue using CBC data," *IEEE Access*, vol. 12, pp. 112355–112367, 2024, doi: 10.1109/ACCESS.2024.3443299.





BIOGRAPHIES OF AUTHORS

Adiba Haque     obtained her Bachelor's degree in Computer Science and Engineering from North South University, Dhaka, Bangladesh. She is currently working at Standard Chartered Bank, Bangladesh. Her research interests include machine learning and networking. She can be contacted at email: adiba.haque@northsouth.edu.







Sanjida Islam     completed her Bachelor's in Computer Science and Engineering from North South University, Dhaka, Bangladesh. Her research interest includes database and machine learning. She can be contacted at email: sanjida.islam16@northsouth.edu.







Nusrat Rahim Mim     finished her Bachelor of Science in Computer Science and Engineering from North South University, Dhaka, Bangladesh. Her research focus areas are networking and machine learning. She can be contacted at email: nusrat.mim@northsouth.edu.







Sabrina Mannan Meem     has completed her Bachelor of Science in Computer Science and Engineering from North South University, Dhaka, Bangladesh. She is working as a trainee engineer specializing in data science in a government grant training program. Her research includes data analytics, machine learning, and natural language processing. Her thesis article on sentimental analysis is published in the Vietnam Journal of Computer Science. She can be contacted at email: sabrina.meem@northsouth.edu.



Ananya Saha     obtained his B.Sc. degree in Computer Science and Engineering from North South University, Dhaka, Bangladesh. His research interests involve artificial intelligence and deep learning. He can be contacted at email: ananya.saha@northsouth.edu.



Riasat Khan     earned his B.Sc. degree in Electrical and Electronic Engineering from the Islamic University of Technology, Bangladesh, in 2010. He further pursued his academic journey, completing both the M.Sc. and Ph.D. degrees in Electrical Engineering at New Mexico State University, Las Cruces, USA, in 2018. Presently, he holds the position of Associate Professor in the Department of Electrical and Computer Engineering at North South University, Dhaka, Bangladesh. His research interests include machine learning, computational bioelectromagnetics, model order reduction, and power electronics. He can be contacted at email: riasat.khan@northsouth.edu.