Predicting Diabetes in Bangladesh Using Machine Learning: A Data-Driven Approach

Tonny Shekha Kar[1], Jothirmoy Sarker Shuvo[2], Md. Robiul Islam Roman[3]
[1]tonny.kar@aiub.edu, [2]22-46473-1@student.aiub.edu, [3]22-46490-1@student.aiub.edu
[1,2]Department of Computer Science, [3]Department of Electrical & Electronic Engineering,
[1,2,3]American International University Bangladesh, Bangladesh

Abstract: Diabetes is a chronic condition caused by inadequate insulin synthesis, which prevents the body from processing blood sugar. Diabetes' etiology is still mostly unknown. Furthermore, diabetes is not routinely measured. Since there is no known treatment for diabetes, it is crucial for those who have disease to monitor their blood sugar levels in order to manage and maintain their health. Therefore, safeguards against diabetes or early detection are necessary. Diabetes symptoms can occur suddenly, and it can be mild so it can take years to notice. Diabetes can damage the heart, eyes, kidneys, nerves and damage blood vessels over time. As of today, no known means to prevent diabetes nor its cause are known so early diagnosis or predicting diabetes is necessary to prevent the worst effects of diabetes. Additionally, People in Bangladesh are still ignorant about diabetes and unaware exactly when diabetes has to be measured. In this paper, we are going to predict diabetes by using machine learning. We compared conventional machine learning with deep learning approaches. For the conventional machine learning method, we considered the most commonly used classifiers: K-Nearest Neighbors (KNN) and Random Forest (RF). On the other hand, for Deep Learning (DL) we employed a fully Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) Algorithm to predict and detect diabetes patients.

Keywords: Artificial Neural Networks, Diabetes Prediction, Deep Learning, Random Forest.

I.Introduction

Diabetes has a significant impact and becomes a global issue worldwide. Bangladesh is among the seven nations that make up the IDF (International Diabetes Federation). By 2050, the IDF predicted there would be 853 million diabetics worldwide, including 184.5 million in the SEA Region. In South-East Asia 374,000 estimated deaths were caused by diabetes in 2024 [1]. As of 2024 the number of adults with diabetes in Bangladesh is around 13.9 million and by 2050 it is predicted to be 23.1 million [1]. In global scale diabetes is infecting significant health and economic burden. There are 529 million patients worldwide with diabetes and the age-standardized diabetes prevalence is 6.1% globally as of 2021 [2]. Type 2 diabetes accounts for 96% of diabetes cases in 2021. By 2025 it is projected that 1.31 billion people will have diabetes [2].

Medical diagnosis of diabetes is one of the most challenging tasks in medicine [3]. Many parameters must be gathered to perform the prediction such as plasma glucose concentration, serum insulin, body mass, and age [4]-[5], and analyzing and making the final decision takes a long time [3]. Consequently, using advanced technologies such as machine learning algorithms are beneficial rather than traditional approaches [6]. This approach can help physicians make more accurate decisions in a short time and it can reduce cost [3].

According to the World Health Organization (WHO), the number of people with diabetes has been continuously rising [7]. It has become alarming in effect in Bangladesh in recent times. It occurs when the body becomes insulin-resistant or cannot produce enough. Diabetes increases the chance of Heart disease, Kidney failure, Nerve damage, Eye problems, Foot ulcers, and infections. Machine learning and deep learning have revolutionized the field of the medical sector. By analyzing the database, we can see the correlation among factors like age, pulse rate, systolic, diastolic, glucose, BMI, and diabetic binary [7]. Through these insights, healthcare professionals can better understand the relationships between these parameters, facilitating early diagnosis, risk prediction, and targeted interventions for conditions like diabetes and cardiovascular diseases. Moreover, the integration of machine learning models in medical systems ensures enhanced decision-making accuracy and improved patient outcomes [3].

Symptoms

Diabetes symptoms can occur suddenly and without any warning. In the case of type 2 diabetes the symptoms can take years to be noticed, and it can be mild [7]. Symptoms of diabetes include feeling very thirsty, blurred vision, feeling tired, losing weight unintentionally, needing to urinate more often than usual. Diabetes can damage the heart, eyes, kidneys, nerves and damage blood vessels over time. Diabetes increases the risk of heart attack, stroke and kidney failure. Diabetes can cause foot ulcers because of nerve damage and poor blood flow [7].

Type 1 diabetes

Type 1 diabetes is known as insulin-dependent, childhood-onset, or juvenile, it is also characterized by insulin production deficiency and requires daily administration of insulin [8]. In 2017 9 million people with type 1 diabetes were found and most of them live in high-income countries [7]. As of today, no known means to prevent diabetes nor its cause are known [8].

Type 2 diabetes

Type 2 diabetes stops the body from using insulin properly and affects how your body uses sugar or glucose for energy, which can lead to high levels of blood sugar if not treated. Over time it can cause serious damage to the body, especially nerves and blood vessels damage are caused by type 2 diabetes [7].

The worst effects of type 2 diabetes can be prevented by early diagnosis. Regular check-ups and blood tests with a healthcare provider are the best ways to detect diabetes early. Symptoms of type 2 diabetes may be similar to type 1 diabetes but often less marked as a result it may be diagnosed several years after complications have already arisen [7].

Recently type 2 diabetes was seen only in adult onset but now it is also occurring in children frequently. Type 2 diabetes was formerly called non-insulin dependent or adult onset and more than 95% of people with diabetes have type 2 diabetes [8].

Gestational diabetes

Hyperglycemia with blood glucose values above normal but that diagnostic of diabetes is gestational diabetes, which occurs during pregnancy. During pregnancy and at delivery women with gestational diabetes are at an increased risk of complications. In the future these women and likely their children are also at risk of type 2 diabetes. Diagnosis of gestational diabetes is done through prenatal screening, rather than reported symptoms.

II.Literature review

Machine learning and deep learning technic in the medical field have become major fields as well for diabetic prediction. Numerous studies have been published in this area. Several studies have taken the Pima Indians Diabetes Dataset (PIDD) [20] as to benchmark dataset for evaluation [9]. Due to their capacity to manage nonlinear relationships in data, studies employing Support Vector Machines (SVM) have demonstrated great accuracy in diabetes prediction [10]. Similarly, research has shown that Random Forest (RF) is effective in handling imbalanced datasets and has been frequently used due to its interpretability and robustness [10]. Also, we used a fully Artificial Neural Network (ANN) for Deep Learning (DL) in order to identify and forecast diabetes patients.

A number of new methods have been created recently. Many contemporary methods have been developed as a result of the advancement of technology for diabetes mellitus diagnosis. The following is a brief overview of the work related to this field. The Pima Indian Diabetes dataset (PIDD) [20] was taken from the UCI machine learning repository database and used by the author in [11]. For five-fold cross-validation, the author employed a Deep Neural Network (DNN) with a prediction accuracy of 98.35%, an F1 score of 98, and an MCC of 97. Moreover, 97.11% accuracy, 96.25% sensitivity, and 98.80% specificity were recorded [11]. KNN and the Naïve Bayes approach were employed by Shetty et al. [12] to predict diabetes. Their method was put into practice as an expert software program that allows users to enter information about patient records and determine if a patient has diabetes or not. Santhanam et al. [13] suggested a method that uses K Means, genetic algorithms, and support vector machines (SVM) to predict the diagnosis of diabetes. The following actions were taken by the system. Update all of the missing values with the mean as the first step. The cleaned dataset is then grouped using K Means to remove extraneous information and outliers, and the best feature is chosen using a genetic algorithm to minimize the features. With SVM, the system's maximum accuracy is 98.82% [13]. In the paper [14], Random Forest, Naive Bayes, and SVM algorithms are used in machine learning experiments on the PIMA dataset to predict diabetes. They performed experiment on Pima Indians Diabetes Database (PIDD) [20] and got the highest accuracy of 76.30%.

The role of Adaboost and Bagging ensemble machine learning approaches is discussed by Sajida et al. in [15]-[16]. Diabetes mellitus and patients are classified as either diabetic or non-diabetic based on diabetes risk variables utilizing the J48 decision tree. The experiment's outcomes demonstrate that Adaboost machine learning ensemble technique performs better than a J48 decision tree in terms of comparatively bagging.

The primary goal of the diabetes prediction system created by Orabi et al. in [17] is to forecast whether a candidate would get diabetes at a given age. By using decision trees, the suggested system is created using the machine learning approach. The results were good since the system that was created was able to forecast the incidence of diabetes at a specific age with greater accuracy by applying decision trees [18]–[19].

From the creation and performance analysis of innovative data mining-based methods for diabetes detection, prediction, and classification to survey and review studies, as demonstrated in [21]-[22], a wide range of literature has contributed to the field of diabetes diagnosis and prediction. Several data mining methods for diabetes detection are examined and addressed in [23], [24], and [25]. Similar to this, a thorough analysis of the use of data mining techniques for diabetes, together with the related data sets, methodologies, software, and technologies, is conducted in [26].

Four machine learning methods are utilized in this experimental investigation [12] to predict early-stage diabetes: Random Forest, K-nearest neighbor, Support Vector Machine, and Linear Discriminant investigation. At 87.66%, the Random Forest classifier has a high accuracy.

To put it another way, the authors of the research [13] have developed models to categorize and forecast problems from diabetes. In this study, eight diabetes complications were predicted and categorized using a variety of supervised classification methods. Metabolic syndrome, dyslipidemia, nephropathy, diabetic foot, obesity, and retinopathy are among the criteria that are among the consequences.

Machine learning algorithms such support vector machines, logistic regression, decision trees, random forests, gradient boost, K-nearest neighbor, and the Naive Bayes algorithm were tested by the authors of this article [15]. The results showed that, in comparison to the other methods, the Naive Bayes and Random Forest classifiers obtained 80% accuracy.

III.Methodology

Brief Description of Algorithms Used:

A.1. Computational Models and Artificial Intelligence

Artificial intelligence models like machine learning and deep learning are the subject of ongoing research due to the rapid growth of massive data and advancements in hardware technology. The most useful feature of deep learning and machine learning is generalization, yet it has proven challenging. The vast amounts of data that have been gathered from the same data distribution are divided into training and test data under the supposition that they are independent and have the same distribution so that current machine learning algorithms can learn and make decisions. The need for machine learning algorithms to be extended so that they perform similarly when learning new problems and data is one of the core difficulties. The majority of machine learning algorithms try to tackle the generalization problem by learning from vast amounts of data that are independent and have the same distribution, or either disregard or exploit pertinent information from the information—such as domain shift and temporal structure as redundant. For this reason, there are certain issues with machine learning that need to be addressed fixed [31].

A.2. Support Vector Machine (SVM):

The typical collection of supervised machine learning models used for classification includes SVM. Finding the optimal highest-margin separation hyperplane between two classes is the goal of a support vector machine given a training sample with two classes [26]. The hyperplane shouldn't be nearer the data points of the other class for better generalization. It is necessary to select a hyperplane that is far from the data points in each category. The support vectors are points that are closest to the margin of the classifier [32].

A.3. Artificial Neural Networks (ANN):

The mechanism of the human nervous system serves as the model for artificial neural networks, or ANNs. In addition to handling fuzzy scenarios, ANN may learn from experience and extract the necessary features from inputs that contain unnecessary information. Three layers make up an ANN's basic architecture: the input, output, and hidden layers. Neurons in the buried layer manipulate data in order to improve learning capacity. The neural network's performance is also impacted by the number of hidden layers; an overfitting issue will arise from having too many hidden layers.[11]

A.4. K-Nearest Neighbors (KNN)

Data are collected from many sources in the modern world in order to enable analysis, form conclusions, and test hypotheses. Missing data in collected data, however, are not uncommon, owing to problems of extraction or human mistakes incurred during data collection. Therefore, in the preprocessing of data addressing missing values is a very significant step. Selection of the correct method to employ in imputing these missing values is of great significance since it can significantly affect model performance. KNN imputer, which can be accessed for free within the sci-kit-learn library, is one of the standard methods employed in imputing missing data. This method is a substitute for traditional methods. The KNN imputer utilizes the Euclidean distance matrix to find the nearest

neighbors and thereby allow imputation of missing values in the observations. In calculating Euclidean distance, it ignores missing values and gives greater weight to non-missing coordinates [33].

A.5. Random Forest (RF)

The robustness and accuracy of Random Forest, a potent supervised classification method developed by Leo Breiman, are well known. There are two phases to it: prediction and forest development. During the forest creation phase, a group of decision trees is constructed. From the total number of features (m), a random feature subset (R), where R is much smaller than m, is chosen for each tree. The method is continued until a predetermined number of nodes (l) is reached, using the best split among these features to divide the data. Several trees are created by repeating this process, creating a "forest" of n trees. All tree output is combined by the model at prediction time, typically by majority vote for categorization. At prediction time, the model aggregates the output of all trees—usually by majority vote for classification or averaging for regression—to produce the final prediction, which increases accuracy and reduces the risk of overfitting [34].

A.6. Logistic regression analysis

To evaluate the statistical relationship between the outcomes of interest (amputation, mortality, and reintervention) and patient attributes such demographics and comorbidities, we construct multivariable logistic regression models. Presented are p-values and model-estimated odds ratios (ORs) at a significance level of 0.05. The absolute value of the t-statistic for each model parameter was used to examine the relative importance of each predictor in the logistic regression models. Since the significance of the parameter estimate is dependent on the t distribution, which is frequently measured by the p-value, the bigger the t-statistic, the more important the predictor in traditional logistic regression. The t-statistic is calculated by dividing the parameter estimate by the standard error [35].

Dataset Used:

The DiaHealth dataset is a thorough compilation of 5,437 patient records that aims to predict Type 2 diabetes and is accessible through the UIU Datasets Repository. This Bangladeshi dataset, which is made available by PKSF and CMED Health Ltd., includes 14 features, including age, gender, blood pressure, glucose, BMI, and medical history. It provides information on diabetes risk variables and is especially useful for health informatics research and predictive model development. The information can be used for categorization in healthcare management and early disease detection [36].

Table.1 Dataset Description.

Database	No. of Attributes	No. of Instances
DiaHealth	14	5437

In addition to its core use in diabetes prediction, data scientists and researchers in the fields of public health and machine learning can benefit greatly from the DiaHealth dataset. By examining clinical and demographic traits including pulse rate, family medical history, history of hypertension, stroke, and cardiovascular disease, researchers can uncover complex correlations and patterns specific to the Bangladeshi community. This helps to improve the creation of regional health models and further aids global research on non-communicable illnesses. Given the range of analytical techniques it provides, including feature selection, model evaluation, and health risk classification, the dataset is a versatile tool for both theoretical and practical healthcare solutions [36].

Table.2 Variable's description of the dataset.

Variables	Description	Data Type
age	Age of the individual in years.	int64
gender	Biological sex of the individual.	Object
pulse_rate	Number of heart-beats per minute.	int64
systolic_bp	Upper value of blood pressure during heartbeats.	int64

diastolic_bp	Lower value of blood pressure between heartbeats.	int64
glucose	Blood sugar level.	float64
height	Height of the individual in centimeters.	float64
weight	Body weight of the individual.	float64
bmi	Body Mass Index, a measure of body fat based on height and weight.	float64
family_diabetes	Indicates if there's a family history of diabetes.	int64
hypertensive	Indicates if the person has high blood pressure.	int64
family_hypertension	Indicates family history of hypertension.	int64
cardiovascular_disease	Presence of heart-related diseases.	int64
Stroke	History of stroke.	int64
diabetic	Indicates have diabetes or not.	object

Methods Used

C.1. Loading And Preprocessing Data

The DiaHealth dataset was imported into a python environment using pandas library. The dataset was loaded in a CSV format that contains the data of 5437 people. The dataset was subjected to a series of preprocessing such as handling missing values (no null value found), data type conversion (gender, diabetic), Normalization or standardization of numerical features where required.

C.1. Train- test Split

This study used a technique called random oversampling to address the dataset's class imbalance, which occurred when there were significantly more non-diabetic samples than diabetes samples. In particular, to equal the size of the majority class (non-diabetic patients), the minority class (diabetic cases) was up-sampled. This was accomplished by developing synthetic examples based on the existing ones by randomly resampling the diabetes occurrences with replacement. By ensuring that the model receives an equal number of instances from both classes during training, the up-sampling procedure lessens bias towards the majority class and enhances the model's capacity to accurately detect cases of diabetes. To guarantee a random distribution of samples prior to model training, the freshly balanced dataset was shuffled following resampling. In binary classification issues, such as those in medical diagnostics, where the cost of incorrectly categorizing the minority class is large, this method is crucial.

The methodology's 80-20 split was used to separate the balanced dataset into training and testing subsets. The goal variable (y_balanced), which is the 'diabetic' column that indicates whether a person has diabetes or not, was initially isolated from the features (X_balanced). The dataset was then randomly partitioned using the train_test_split method from the sklearn.model_selection module. In particular, 80% of the data was used to train the machine learning models, with the remaining 20% set aside to assess how well they performed on data that had not yet been seen. For results to be repeatable, a fixed random_state was set. This stage is essential for confirming the models' capacity for generalization and avoiding overfitting.

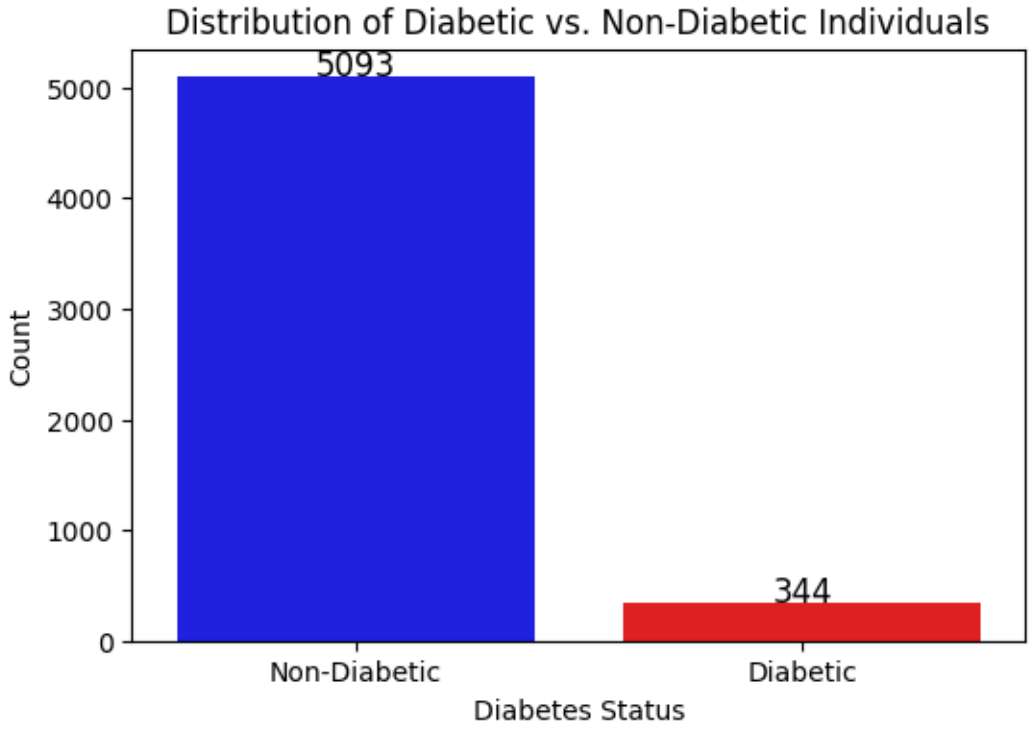

Fig.1 Distribution of Diabetic vs. Non-Diabetic Individuals.

Variable	Minimum	Maximum	Mean	Std. Deviation
age	8	112	45.53	14.321
pulse_rate	5	133	76.77	12.29
systolic_bp	62	231	133.86	22.293
diastolic_bp	45	119	82.06	12.49
glucose	0	33.46	7.5407	2.92308
bmi	1.22	574.13	22.4723	8.77876

Table 3: Statistical Datum of DiaHealth Dataset.

With a total value of 5437, the dataset exhibits notable variation across all variables. There is a large age range from 8 to 112 years old, with an average age of 45.53 years and a standard deviation of 14.32. The pulse rate's standard deviation is 12.29 and its mean is 76.77 beats per minute. The average diastolic blood pressure is 82.06 mmHg with a standard deviation of 12.49 and the average systolic blood pressure is 133.86 mmHg with a standard deviation of 22.29. BMI averages 22.47 and glucose averages 7.54, both of which exhibit significant variability (standard deviations of 8.78 and 2.92, respectively). In terms of health indicators, these figures point to a varied population.

The Pearson correlation coefficients are shown in this correlation table for the following variables: age, pulse rate, glucose, BMI, systolic and diastolic blood pressure (BP), and diabetes status. Significant correlations (at the 0.01 or 0.05 level) are indicated by asterisks. For example, age is positively correlated with systolic blood pressure (r = 0.367) and glucose (r = 0.107), but negatively correlated with BMI (r = -0.093). Systolic and diastolic blood pressure have a strong correlation (r = 0.716), although glucose and diabetes status have a modestly positive relationship (r = 0.554). These correlations, which illustrate the relationships between variables, may facilitate the identification of risk factors or health trends.

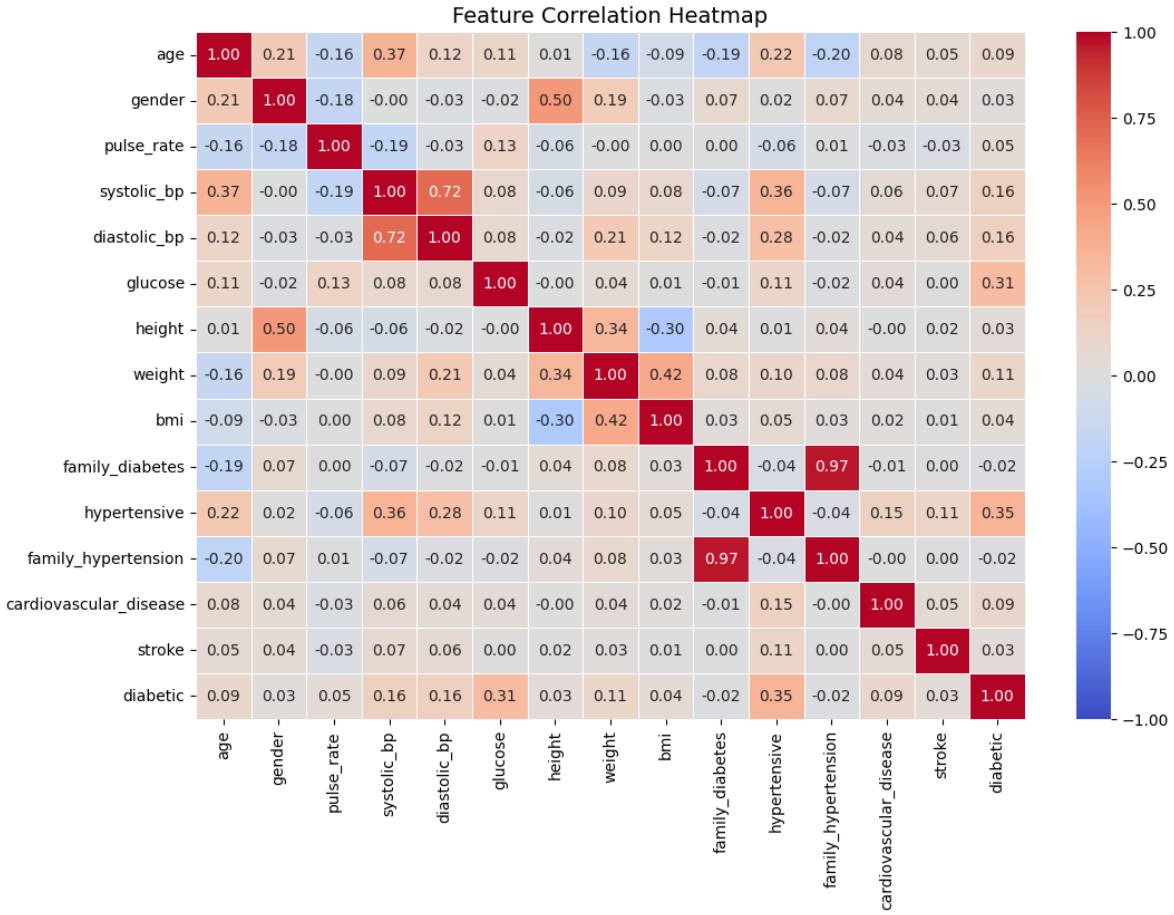

Fig 2. Heat Map.

IV.results

Using Accuracy and F1 Score, the table 4 compares the performance of six distinct machine learning models for diabetes prediction. With an accuracy of 0.9887 and an F1 Score of 0.9888, the Random Forest classifier outperformed all the other models, demonstrating its superior capacity to accurately classify both diabetic and non-diabetic individuals. K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANN) also shown exceptionally strong performance; KNN came in second with values of 0.9372 and 0.9405, respectively, while ANN achieved an accuracy of 0.9524 and an F1 Score of 0.9539. These findings imply that deep learning and ensemble approaches are both quite successful in completing this classification assignment.

Model	Accuracy	F1 Score
Support Vector Machine	0.7772	0.7511
Logistic Regression	0.788518	0.775871
Long Short-Term Memory	0.816487	0.811111
K-Nearest Neighbors	0.937193	0.940520
Artificial Neural Networks	0.9524	0.9539
Random Forest	0.988714	0.988786

Table 4: Obtained performance of the model.

On the other hand, conventional models such as Long Short-Term Memory (LSTM), Support Vector Machine (SVM), and Logistic Regression produced poorer performance scores; their F1 Scores were below 0.78 and their accuracies were below 0.79. With an accuracy of 0.8165 and an F1 Score of 0.8111, LSTM fared somewhat better. All things

considered, the findings unequivocally show that models like Random Forest and ANN that can recognize intricate patterns are more accurate in predicting diabetes in unbalanced medical datasets.

V.Conclusion

By comparing the effectiveness of many machine learning models on a publicly available healthcare dataset for diabetes prediction, this work investigates a comparatively understudied topic in the context of Bangladesh. With 5,093 (93.67%) of the 5,437 people in the dataset having diabetes and just 344 (6.33%) not having the disease, there is a clear class disparity. Oversampling techniques were used to balance the dataset in order to solve this, preventing problems such as overfitting and model bias. The study admits that the original dataset's imbalance and relatively small size are significant limitations that may have an impact on the results' generalizability, even if the majority of models attained excellent accuracy. To improve the findings' validity and application, future study using larger and more varied datasets is advised.

References

[1] Magliano, D. J., Boyko, E. J., & International Diabetes Federation. (2025). IDF Diabetes Atlas 11th Edition - 2025. In Optima (Ed.), IDF Diabetes Atlas (11th ed., p. 84–85) [Report]. https://diabetesatlas.org.

[2] GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet. 2023 Jul 15;402(10397):203-234. doi: 10.1016/S0140-6736(23)01301-6. Epub 2023 Jun 22. Erratum in: Lancet. 2023 Sep 30;402(10408):1132. doi: 10.1016/S0140-6736(23)02044-5. Erratum in: Lancet. 2025 Jan 18;405(10474):202. doi: 10.1016/S0140-6736(25)00053-4. PMID: 37356446; PMCID: PMC10364581.

[3] S. Islam Ayon, Md. M. Islam, and M. Milon Islam, “Diabetes prediction: a deep learning approach,” International Journal of Information Engineering and Electronic Business, vol. 11, no. 2, pp. 21–27, 2019.

[4] F. A. Khan, K. Zeb, M. AL-Rakhami, A. Derhab, and S. A. C. Bukhari, “Detection and prediction of diabetes using data mining: a comprehensive review,” IEEE Access, vol. 9, pp. 43711–43735, 2021.

[5] American Diabetes Association, “Diagnosis and classification of diabetes mellitus,” Diabetes Care, vol. 37, no. 1, pp. 81–90, 2014.

[6] M. Djerioui, Y. Brik, M. Ladjal, and B. Attallah, “Neighborhood component analysis and support vector machines for heart disease prediction,” Ing´enierie des Syst`emes d’Information, vol. 24, no. 6, pp. 591–595, 2019.

[7] World Health Organization: WHO & World Health Organization: WHO. (2024, November 14). Diabetes. Retrieved April 17, 2025, from https://www.who.int/news-room/fact-sheets/detail/diabetes.

[8] Roglic, Gojka. WHO Global report on diabetes: A summary. International Journal of Noncommunicable Diseases 1(1):p 3-8, Apr–Jun 2016. | DOI: 10.4103/2468-8827.184853.

[9] N. Nisha Nadhira Nazirun et al., "Prediction Models for Type 2 Diabetes Progression: A Systematic Review," in IEEE Access, vol. 12, pp. 161595-161619, 2024, doi: 10.1109/ACCESS.2024.3432118.

[10] F. A. Khan, K. Zeb, M. Al-Rakhami, A. Derhab and S. A. C. Bukhari, "Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review," in IEEE Access, vol. 9, pp. 43711-43735, 2021, doi: 10.1109/ACCESS.2021.3059343.

[11] Ayon, Safial & Islam, Md. (2019). Diabetes Prediction: A Deep Learning Approach. International Journal of Information Engineering and Electronic Business. 11. 21-27. 10.5815/ijieeb.2019.02.03.

[12] Shetty D, Rit K, Shaikh S, Patil N. Diabetes disease prediction using data mining. Innovations in information, embedded and communication systems (ICIIECS), 2017 international conference on. 2017. p. 1–5.

[13] T. Santhanam and M.S Padmavathi, “Application of K Means and Genetic Algorithms for Dimension Reduction by Integrating SNM for Diabetes Diagnosis,” Procedia Computer Science, vol. 47, pp. 76-83, 2015.

[14] D. Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms,” Procedia Computer Science, vol. 132, pp. 1578–1585, 2018, https://doi.org/10.1016/j.procs.2018.05.122.

[15] Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science 82, 115–121. doi:10.1016/j.procs.2016.04.016.

[16] Nai-Arun, N., Sittidech, P., 2014. Ensemble Learning Model for Diabetes Classification. Advanced Materials Research 931-932, 1427–1431. Doi:10.4028/www.scientific.net/AMR.931-932.1427.

[17] Orabi, K.M., Kamal, Y.M., Rabah, T.M., 2016. Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer. Springer. pp. 420–427.

[18] Priyam, A., Gupta, R., Rathee, A., Srivastava, S., 2013. Comparative Analysis of Decision Tree Classification Algorithms. International JournalofCurrentEngineeringandTechnologyVol.3,334–337. doi:JUNE2013,arXiv:ISSN 2277- 4106.

[19] Esposito, F., Malerba, D., Semeraro, G., Kay, J., 1997. A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 476–491. doi:10.1109/34.589207.

[20] Kayaer, K., Tulay, 2003. Medical diagnosis on Pima Indian diabetes using general regression neural networks, in: Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP), pp. 181–184.

[21] B. L. Shivakumar and S. Alby, "A survey on data-mining technologies for prediction and diagnosis of diabetes, in Proc. Int. Conf. Intell. Comput. Appl., Mar. 2014, pp. 167173.

[22] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, Machine learning and data mining methods in diabetes research, Comput.Struct.Biotechnol.J., vol. 15, pp. 104116, Jan.2017.

[23] K. Rajalakshmi and D. S. S. Dhenakaran, Analysis of data mining prediction techniques in healthcare management system, Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 5, no. 4, pp. 13431347, Apr. 2015.

[24] G. Visalatchi, S. J. Gnanasoundhari, and M. Balamurugan, A survey on data mining methods and techniques for diabetes mellitus, Int. J. Comput. Sci. Mobile Appl., vol. 2, no. 2, pp. 100105, 2014.

[25] D. Tomar and S. Agarwal, A survey on data mining approaches for healthcare, Int. J. Bio-Sci. Bio-Technol., vol. 5, no. 5, pp. 241266, Oct. 2013.

[26] M. Marinov, A. S. M. Mosa, I. Yoo, and S. A. Boren, Data-mining technologies for diabetes: Asystematicreview, J.DiabetesSci.Technol., vol. 5, no. 6, pp. 15491556, Nov. 2011.

[27] G. Tripathi and R. Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning," in 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Jun. 2020, pp. 1009–1014. doi:10.1109/ICRITO48877.2020 .9197832.

[28] Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, "A Machine Learning Approach to Predicting Diabetes Complications," Healthcare, vol. 9, no. 12, Art. no. 12, Dec. 2021, doi: 10.3390/healthcare9121712.

[29] K. Pavani, P. Anjaiah, N. V. Krishna Rao, Y. Deepthi, D. Noel, and V. Lokesh, "Diabetes Prediction Using Machine Learning Techniques: AComparative Analysis," in Energy Systems, Drives and Automations, Singapore, 2020, pp. 419–428. doi: 10.1007/978-981-15-5089-8 41.

[30] Schölkopf, B. Causality for machine learning. In Probabilistic and Causal Inference: The Works of Judea Pearl; Geffner, H., Ed.; Association for Computing Machinery: New York, NY, USA, 2022; pp. 765–804.

[31] Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: Proceeding soft The Second International Conference on Soft Computing for Problem Solving (SocProS2012), December28-30,2012, Springer.pp.1027–1038.

[32] Greeshma, U.; Annalakshmi, S. Artificial Neural Network (Research paper on basics of ANN). Int. J. Sci. Eng. Res. 2015, 110–115.

[33] K. Kannadasan, D. R. Edla and V. Kuppili, "Type 2 diabetes data classification using stacked autoencoders in deep neural networks", Clin. Epidemiol. Global Health, vol. 7, no. 4, pp. 530-535, Dec. 2019.

[34] Random Forest and Decision Trees, By Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran, Maqsood, Computer Engineer UUET Peshwa, Pakisthan.

[35] Kirasich K, Smith T, Sader B. Random forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Sci Rev. 2018;1(3):9.

[36] UIU Datasets Repository System. (n.d.). Retrieved April 25, 2025, from https://data.uiu.ac.bd/dataset/iriic/1.