Advancing Diabetes Prediction: A Comprehensive Study Integrating Deep Learning and Machine Learning Approaches

1st Mohammed Z Waughfa
Department of CSE
Ahsanullah University of Science Technology
Dhaka, Bangladesh
190204037@aust.edu

2nd Sumaiya Siddiqua Mumu
Department of CSE
Ahsanullah University of Science Technology
Dhaka, Bangladesh
190204040@aust.edu

3rd Syeda Samia Sultana
Department of CSE
Ahsanullah University of Science Technology
Dhaka, Bangladesh
190204048@aust.edu

4th Imranul Islam Adnan
Department of CSE
Ahsanullah University of Science Technology
Dhaka, Bangladesh
190204053@aust.edu

***Abstract*— Diabetes is one of the major chronic diseases of today's world. A huge number of people are suffering from this disease. While there is no cure for diabetes, early prevention, and treatment help patients to lead a healthy life. Recently, for predicting diabetes machine learning and deep learning approaches have become popular. Utilizing the Pima Indian dataset, this study investigated machine learning and deep learning techniques for diabetes prediction. A comprehensive evaluation was conducted on the performance of various algorithms, including Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest, K-Nearest Neighbors (KNN), and Decision Tree, employing criteria such as accuracy, precision, and recall. Notably, the results demonstrated the superiority of MLP, achieving an accuracy rate of 85%. This underscored MLP's efficacy in early-stage diabetes recognition and management. The research emphasized the significance of machine learning in the healthcare industry and advocated for ongoing research to enhance diabetes prediction algorithms.**

Index Terms*— *Diabetic Prediction, SVM, MLP ,Random Forest, KNN and decision tree algorithms

I. Introduction

Diabetes is a chronic disease that occurs due to a lack of insulin or lack of efficient use of insulin in the human body. [1] Diabetes has 2 types: type 1 and type 2 diabetes. Among these 2 types, majority of diabetes patients are suffering from type 2 diabetes. [2] Diabetes affects a large number of people worldwide and is rapidly increasing in prevalence. It is estimated by International Diabetes Federation that over 285 million people are suffering from diabetes today, and that number will rise to 380 million in the next 20 years. [3]

Recently machine learning and deep learning based approach has become popular when it comes to diabetes prediction. These models enable proactive management and lifestyle modifications by identifying hidden risk factors through the analysis of different patient data. The prevention of complications is aided by early detection, and the management of diabetes is made more efficient overall through tailored care that maximizes healthcare resources.

The study involves a thorough investigation, using techniques such as Principal Component investigation, to uncover critical traits for machine learning. The emphasis was on improving the model for identifying diabetes. The project concluded with the deployment of three different machine learning classification algorithms: Decision Tree, Support Vector Machine (SVM), and Random Forest. These algorithms were selected to improve the accuracy and efficiency of the diabetes detection procedure in the experiment.

II. Related Work

Deepti Sisodia et al. [4] proposed using Naive Bayes, Decision Tree, and Naive Bayes, to determine the likelihood of diabetes in patients with maximum accuracy. Accuracy, Recall, Precision, and F-Measure were utilized to evaluate the performances of the three models. The accuracy of each algorithm is evaluated based on both correct and incorrect instances. The research found that Naive Bayes achieves 76.30% accuracy which is the highest. Conversely, SVM and Decision Tree gives an accuracy of 65.10% and 73.82% respectively.

In another study, Subhash Chandra Gupta et al. [5] presented a disease prediction model for classifying diabetic patients using machine learning classification algorithms. Decision Tree, KNN, SVM, and Random Forest algorithms were applied to create four models. The hyperparameters of the classifiers were tuned to improve their performance. The best prediction model achieved the highest F1score of 75.68% with 88.61% accuracy. Among the four models, on the dataset model D3, the Random Forest classifier performed the best. Chaitanya Sonawane et al. [6] proposed using SVM and Decision Tree to determine which approach is better for predicting diabetes. Pima Indian diabetes dataset was used for this research. Accuracy, precision, and recall these three evaluation metrics were used to compare the performance of the models. SVM algorithm achieved an accuracy of 76.6% which is better than the decision tree algorithm, which achieved 75% accuracy.

Nour Abdulhadi et al. [7] utilized supervised learning methods to build a model for diabetes detection. The paper

presented multiple techniques and models, including Logistic Regression, Linear Discriminant Analysis (LDA), and Linear Support Vector Machine, with accuracies ranging from 79% to 82%. The Random Forest Classifier achieved the highest accuracy of 82%. The dataset used in the research consisted of 768 instances with 9 attributes, including the target variable. The data was preprocessed, missing values were filled, and the dataset was split into a training set and a test set. The models were trained on the training set and evaluated on the test set to assess their ability to generalize to new data.

III. Dataset

A. Data Collection

The collected dataset originated from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset describes whether a patient is diabetic or not based on criteria included in the dataset. All the patients described in the dataset are Indian women over the age of 21. [8] The dataset consists of 768 instances and 9 attributes, including the target variable 'Outcome' which states the diabetic condition of the patients. Table I represents the description of the attributes of the dataset.

TABLE I
DESCRIPTION OF ATTRIBUTES OF THE DATASET

Attributes	Range	Description
Pregnancies	0-17	Number of times pregnant
Glucose	0-199	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Blood Pressure	0-122	Diastolic blood pressure (mm Hg)
BMI	0-67.1	Body mass index = (weight in kg/(height in m)^2)
Skin Thickness	0-99	Triceps skin fold thickness (mm)
Diabetes Pedigree Function	0.078-2.42	A function that scores the likelihood of diabetes based on family history
Age	21-81	Age in years
Insulin	0-846	2-Hour serum insulin (mu U/ml)
Outcome	0-1	Class variable, diagnoses classes: 0 = healthy, 1 = diagnosed with diabetes

B. Data Preprocessing

For model training, the quality of the dataset must be ensured. Real-world data often contains missing values, errors, and outliers as well. The dataset quality significantly affects the result of a model. Hence, preprocessing of data is very important. It helps to minimize the effects of errors and increases the accuracy of the model.

There were a few instances in the collected dataset where there were multiple null values. For instance, there have been cases where the blood pressure value was zero which is illogical. Instead of removing instances with zeros since there were so few of them (768), the values were filled in using the mean. The correlation between features shown in Figure 3 was generated after preprocessing the original dataset.

C. Correlation Between Features

Correlation values were computed to determine the extent to which an attribute influences the target attribute (Outcome) or whether it impacts other characteristics in order to gain additional insight into the data. Correlation values were calculated by using Pearson correlation method. It determines the measure of the linear relationship between those two features by multiplying the covariance of each feature by the product of its standard deviations.

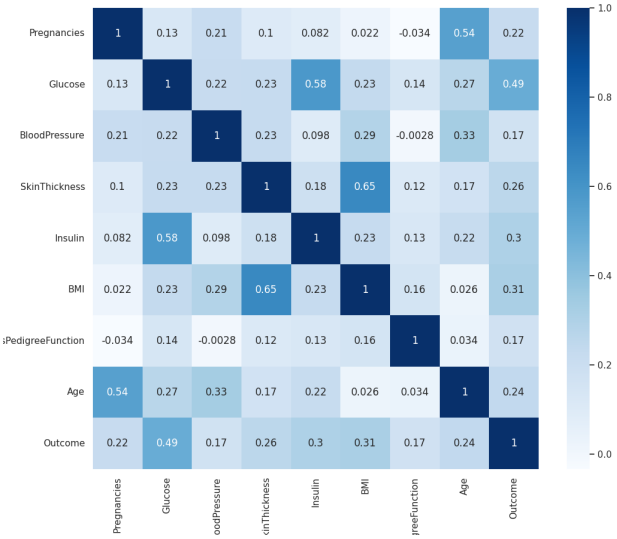

Fig. 1. The correlation represented by Heat map

IV. METHODOLOGY

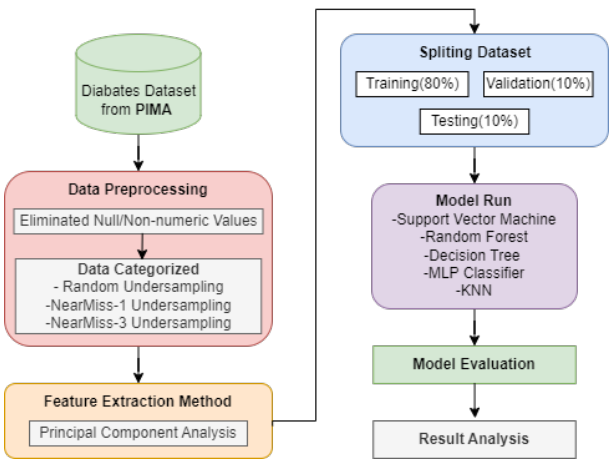

Fig. 2. Proposed Methodology

The methodology used in this study takes a systematic approach to diabetes prediction. The first step entails choosing and obtaining the dataset, which forms the basis for the studies that follow. After processing the dataset, it is compared to the original dataset to see the improvements. To even out

the distribution of classes, three undersampling techniques: NearMiss-1, NearMiss-3, and random undersampling are used. After a thorough evaluation of these undersampling techniques' efficiency, the most promising strategy is chosen for additional analysis.

Principal Component Analysis (PCA) and correlation heat maps are two data visualization approaches used to obtain deeper insights into the dataset.Six PCA components was found to have higher explained variance with lowest error rate and highest accuracy. By helping to comprehend the relationships among the data, these visualizations enable more informed judgments to be made during the modeling process. Three machine learning models—Decision Tree, Support Vector Machine (SVM), Random Forest and one deep learning model— Multilayer Perceptron Classifier (MLP) are put into practice after this exploratory stage.

For KNN, the K value is looped from 1 to 30 to find the best K value for the dataset.Then, in the analysis, it was found that the K-Nearest Neighbors (KNN) algorithm provided the most favorable results, as determined by the ROC curve. This suggests that, in the past investigation, KNN demonstrated the best performance in predicting diabetes based on the evaluation of the Receiver Operating Characteristic (ROC) curve.

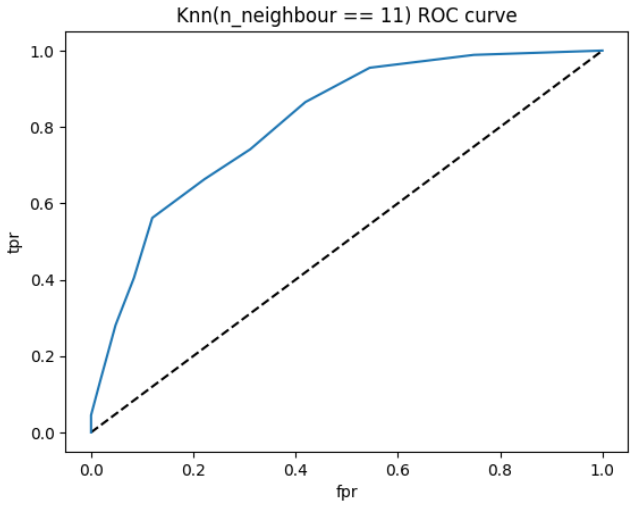

Fig. 3. ROC curve for KNN

Figure 2 provides a visual representation of the Receiver Operating Characteristic (ROC) curves, revealing a consistent observation across all curves. Notably, it is evident that the K-Nearest Neighbors algorithm with 11 neighbors consistently yields the largest area under the curve. This consistent trend underscores the robustness and efficacy of the model at this specific parameter configuration, indicating superior discriminatory power in distinguishing between positive and negative instances. The discernible pattern in the ROC curves underscores the optimal performance achieved by the K-Nearest Neighbors algorithm when configured with 11 neighbors, thereby emphasizing its potential as a favorable choice for diabetes prediction within the studied dataset.

Key performance measures are used to train and assess the models. A thorough comparison of all the models' output is done to determine which one performs the best overall. This detailed comparative analysis is critical for establishing the best effective diabetes prediction model in the context of this investigation. The below figure shows the methodology flow diagram followed in this study.

V. Experimental Results

A. Performance Comparison

Following data preprocessing, five distinct models, namely Decision Tree, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbor(KNN), and Multilayer Perceptron Classifier (MLP) were employed in our analysis. Each model underwent rigorous evaluation to assess its performance and suitability for the given dataset. This strategic approach allowed for a comprehensive exploration of diverse modeling techniques, enhancing the robustness of our analyses.

TABLE II
Model Performance

Model	Class	Precision	Recall	F1-score	Accuracy
SVM	No Diabetes	0.76	0.76	0.76	0.81
	Diabetes	0.85	0.85	0.85	
KNN	No Diabetes	0.79	0.88	0.83	0.77
	Diabetes	0.71	0.56	0.63	
Decison Tree	No Diabetes	0.58	0.67	0.62	0.69
	Diabetes	0.77	0.7	0.73	
Random Forest	No Diabetes	0.71	0.91	0.80	0.80
	Diabetes	0.90	0.70	0.79	
MLP	No Diabetes	0.84	0.76	0.80	0.85
	Diabetes	0.86	0.91	0.88	

B. Result Analysis

The performance of classification algorithms can be compared across different metrics as shown in Table II.

1) **Accuracy**
Accuracy is a metric that calculates the proportion of correctly predicted values out of the total number of instances evaluated.

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

2) **Recall**
Recall, also known as sensitivity, quantifies the proportion of positive values that are accurately classified.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

3) **Precision**
Precision assesses the accuracy of positive predictions within the predicted values of the positive class.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

4) **F1 Score**
The F1-Score quantifies the balanced performance of a classifier by taking into account both the recall and precision rates through their harmonic average.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

In the analysis of five predictive models—Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP)—for diabetes prediction, the results highlight the MLP as the standout performer. With an impressive accuracy of 85%, the MLP model demonstrates robust overall performance. The SVM also showcases strong performance, achieving an accuracy of 81%, indicating its reliability. While the other models perform moderately, the MLP excels in accurately predicting both "No Diabetes" and "Diabetes".

In summary, for a dependable model to predict both "No Diabetes" and "Diabetes," the Multilayer Perceptron stands out as the most effective choice, as supported by the presented metrics.

VI. Conclusion

Diabetes is one of the most common chronic illnesses in modern civilization, impacting a significant proportion of the world's population. Even if there isn't a 100% effective treatment, people can enjoy better lives because to early preventive and treatment methods. Deep learning and machine learning techniques have recently become more popular for diabetes prediction. Using machine learning and deep learning techniques, this work explored the field of diabetes prediction with a specific emphasis on the Pima Indian dataset. After a thorough analysis, the effectiveness of many algorithms—including Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest, K-Nearest Neighbors (KNN), and Decision Tree—was evaluated. Evaluation criteria included important parameters including recall, accuracy, and precision.

Notably, the results demonstrated MLP's outstanding performance, with an accuracy rate of 85%. This demonstrated how effective MLP is in the early detection and treatment of diabetes. The study promoted ongoing research efforts to improve and optimize diabetes prediction algorithms and stressed the critical role that machine learning plays in the healthcare industry. The study's findings highlight how machine learning has the potential to revolutionize diabetes diagnosis and treatment in the future.

References

[1] World Health Organization. (2023). *Diabetes*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] Ayon, Safial Islam, and Md Milon Islam. "Diabetes prediction: a deep learning approach." International Journal of Information Engineering and Electronic Business 12.2 (2019): 21.

[3] Kumari, V. Anuja, and R. Chitra. "Classification of diabetes disease using support vector machine." International Journal of Engineering Research and Applications 3.2 (2013): 1797-1801.

[4] Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, **132**, 1578-1585. International Conference on Computational Intelligence and Data Science. ISSN: 1877-0509.

[5] Gupta, S. C., & Goel, N. (Year). Predictive Modeling and Analytics for Diabetes using Hyperparameter tuned Machine Learning Techniques. *Procedia Computer Science*, **218**, 1257-1269

[6] Sonawane, C., Somwanshi, K., Patil, R., Raut, R. (2023). Diabetic Prediction Using Machine Algorithm SVM and Decision Tree. In *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)*. IEEE. ISBN: 979-8-3503-2379-5.

[7] Nour Abdulhadi, Amman, Jordan ; Amjed Al-Mousa. (2021). Diabetes Detection Using Machine Learning Classification Methods. Available: https://ieeexplore.ieee.org/abstract/document/9491788

[8] Akshay Dattatray Khare. (2022, December). Diabetes Dataset, Version 1. Retrieved December 30, 2023 from https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset