Early detection of diabetes using a supervised classifications algorithm

Darshan Vasoya
Department of Computer Engineering, Marwadi University
Rajkot, India
vasoyadarshan111@gmail.com

Raj Mungara
Department of Computer Engineering, Marwadi University
Rajkot, India
rajmungara1575@gmail.com

Rupesh Rudakiya
Department of Computer Engineering, Marwadi University
Rajkot, India
rkrudakiya2004@gmail.com

Prof. Dhara Joshi
Department of Computer Engineering, Marwadi University
Rajkot, India
dhara.joshi@marwadieducation.edu.in

***00*and it is considered as one of the deadliest and most chronic disease. If it is untreated or unidentified there will be a chances of occurring many complications. The rise in machine learning approaches solves this critical problem. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore, Machine learning classification algorithms namely Logistic Regression, Decision tree and SVM are used in this experiment to detect diabetes at an early stage. Experiments are performed on Real Diabetes Dataset (RDD) which is sourced from Kaggle machine learning repository.The performances of these algorithms are evaluated on various measures like Precision, Accuracy, F-Score, and Recall.**

***Keywords—Diabetes prediction, Early Detection of Diabetes, SVM, K-NN,* Healthcare, Analytics**

I. INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder that continues to grow as a serious global health concern. The International Diabetes Federation (IDF) reported that around 589 million adults were affected by diabetes in 2024, and this figure is expected to increase to nearly 853 million by 2050 [1]. Similarly, the World Health Organization (WHO) estimated that by 2022, approximately 830 million individuals were living with the disease, with the highest burden falling on low- and middle-income nations [2]. These alarming statistics emphasize the importance of early diagnosis and intervention to minimize health complications and mortality.

In recent years, supervised classification machine learning (ML) Algorithms have shown strong potential in predicting diabetes at its early stages. Classical algorithms such as logistic regression**,** decision trees, and support vector machine (SVM) have been widely applied to structured datasets, including the Pima Indians Diabetes Dataset and the Early-Stage Diabetes Risk Prediction Dataset [3], [4]. More advanced techniques, such as deep neural networks (DNNs)**,** have further improved predictive accuracy in many studies [5]. To enhance transparency and clinical trust, explainability frameworks such as SHAP and LIME are increasingly being integrated into diabetes prediction models [6].

Despite these advancements, several challenges persist. Many datasets suffer from class imbalance and are often limited to specific populations, which reduces the generalizability of models [7]. Furthermore, data privacy issues pose barriers to multi-institutional research. To address this, approaches such as federated learning have been explored, allowing institutions to collaboratively train models without directly sharing patient records [8]. These innovations suggest that while supervised ML holds strong promise for early diabetes detection, further work is needed to ensure fairness, reliability, and real-world applicability.

II. LITERATURE REVIEW/SURVEY

In [6], the 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 (DM) 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, the 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 logistic 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.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 the PIDD dataset, their study achieved 94%, 79%, and 69% prediction rates. Notably, Logistic Regression outperformed other algorithms in diabetes prediction efficiency.

III. Proposed Methodolody

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 deadly 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 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 are 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

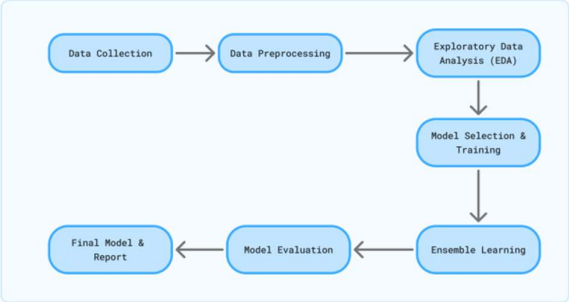

Fig. 1. Proposed model workflow

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], and Logistic Regression [20]. 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 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].

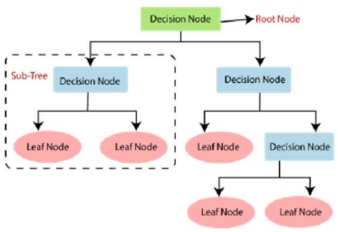

Fig.1. Decision tree

2) 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 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].

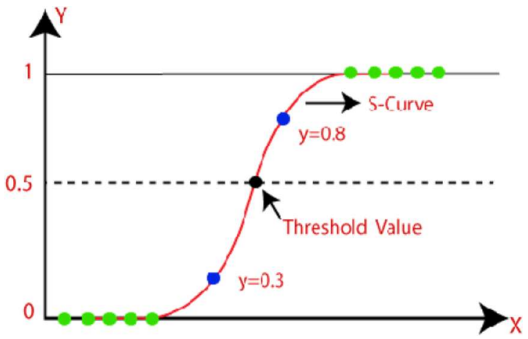

Fig.2. Logistic Regression graph

3) SUPPORT VECTOR MACHINE(SVM)

Support Vector Machine (SVM) is a popular machine learning algorithm used for classification, regression, and outlier detection. The main objective of SVM is to find the hyperplane that separates the data into different classes in the best possible way. In a binary classification problem, SVM algorithm creates a boundary between the two classes by maximizing the margin or distance between the closest data points of each class. The data points closest to the boundary are called support vectors. SVM can also handle multi-class classification problems by using one- vs-all or one-vs-one strategies.

SVM can handle both linear and nonlinear datasets by using different kernel functions such as linear, polynomial, radial basis function (RBF), and sigmoid. The kernel function transforms the data into a higher dimensional space where it is easier to separate the classes. Support Vector Machine (SVM) is a popular machine learning algorithm used for classification, regression, and outlier detection. The main objective of SVM is to find the hyperplane that separates the data into different classes in the best possible way. In a binary classification problem, SVM algorithm creates a boundary between the two classes by maximizing the margin or distance between the closest data points of each

class. The data points closest to the boundary are called support vectors. SVM can also handle multi-class classification problems by using one- vs-all or one-vs-one strategies.

SVM can handle both linear and nonlinear datasets by using different kernel functions such as linear, polynomial, radial basis function (RBF), and sigmoid. The kernel function transforms the data into a higher dimensional space where it is easier to separate the classes.

SVM has several advantages such as:

- It can handle high-dimensional datasets and large sample sizes efficiently.
- It is effective in cases where the number of features is larger than the number of samples.
- It can handle both linear and nonlinear datasets using kernel functions.
- It has a regularization parameter that helps prevent overfitting.

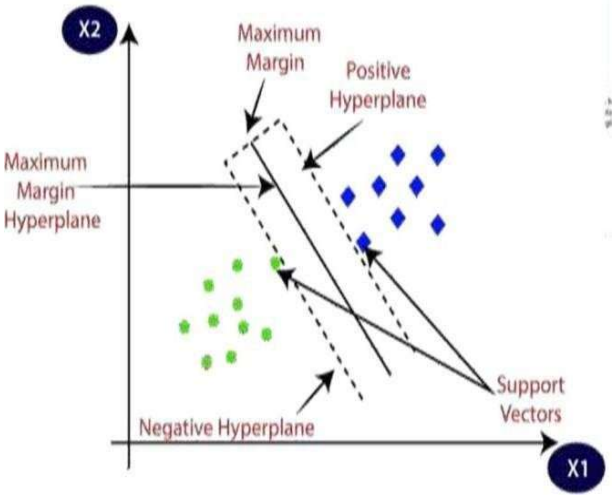

Fig 4 Support Vector Machine

4) K-NEAREST NEIGHBOUR

The k-nearest neighbors (KNN) algorithm is a popular machine learning technique used for classification and regression. It works by finding the k closest data points in the training set to a given input data point and using their labels to make a prediction. To measure the distance between data points, various distance metrics can be used, such as Euclidean distance, Manhattan distance, and Minkowski distance. The optimal value of k can be determined through techniques such as cross-validation, and the choice of k can affect the bias-variance trade-off. KNN has both strengths and weaknesses. Its simplicity and versatility make it easy to understand and apply to a wide range of problems, but it also requires a large amount of training data and can be computationally complex during inference. Variations of KNN include weighted KNN, which gives more weight to closer neighbors, and KNN with kernel functions, which applies a kernel function to the distance metric. Implementing KNN involves preparing and preprocessing the data, calculating distances between data points, and making predictions using the algorithm. Examples of real-world applications of KNN include image classification, sentiment analysis, and personalized recommendations. By understanding the various aspects of KNN and its applications, one can effectively use the algorithm in practice.

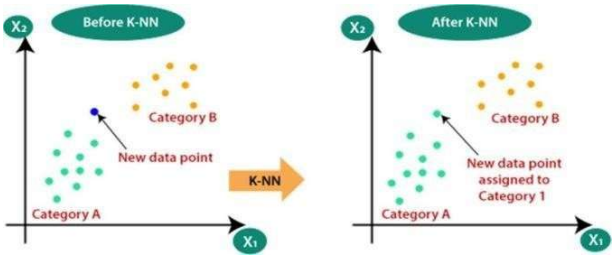

Fig 5 K-Nearest neighbour

IV. RESULTS AND DISCUSSION

Increasing the complexity of the model can help reduce 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, the 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 F1-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 for accuracy [15].

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

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].

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

The success of the algorithm in identifying individuals with cancer is quantified by a metric called "recall" [20]. Mathematically, recall is represented by Equation 3.

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

The term "harmonic mean" describes this method since it balances accuracy and memory. A version of the

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 can work. They both achieved 96% accuracy. On the other hand, GNB didn't work well among these 5 algorithms. GNB achieved 88% model accuracy. In terms of F1-score, DT performs well here. DT achieved 0.98 scores, which is the best score.

TABLE II. CLASSIFICATION REPORT OF THIS MODE

Algorithms	Precisions	Recall	Accuracy
DT + K-NN	97.518	97.518	98.51
SVM	91.6667	91.6667	91.67
LR	94.8718	94.8718	94.87

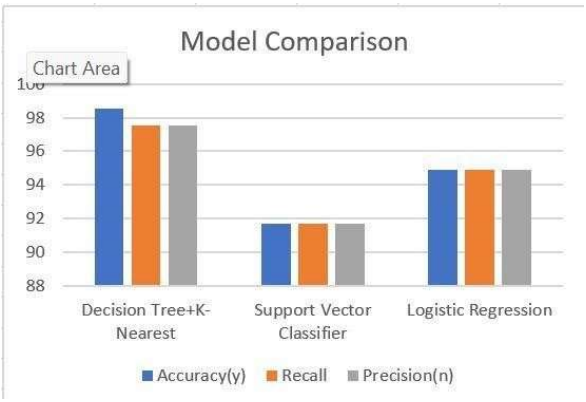

Fig 6: Model Comparison

	precision	recall	f1-score	support
0	0.92	0.92	0.92	52
1	0.96	0.96	0.96	104
accuracy			0.95	156
macro avg	0.94	0.94	0.94	156
weighted avg	0.95	0.95	0.95	156

Fig 7: Classification report of LR

	precision	recall	f1-score	support
0	0.91	0.96	0.93	52
1	0.98	0.95	0.97	104
accuracy			0.96	156
macro avg	0.94	0.96	0.95	156
weighted avg	0.96	0.96	0.96	156

Fig 8: Classification report of DT

	precision	recall	f1-score	support
0	0.85	0.90	0.88	52
1	0.95	0.92	0.94	104
accuracy			0.92	156
macro avg	0.90	0.91	0.91	156
weighted avg	0.92	0.92	0.92	156

Fig 9: Classification Report of SVM

	precision	recall	f1-score	support
0	0.70	0.94	0.80	52
1	0.97	0.80	0.87	104
accuracy			0.85	156
macro avg	0.83	0.87	0.84	156
weighted avg	0.88	0.85	0.85	156

Fig 10: Classification Report of K-NN

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 diabetes. 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. 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. CONCLUSION

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 to develop 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 a better deep learning model.
- Try to find better execution of accuracy.

REFERENCES

[1] T. Islam, T. R. Mona, M. R. Sadik *et al.*, "Early-Stage Diabetes Risk Prediction Using Supervised Machine Learning Algorithms," *Journal of Engineering Research and Reports*, vol. 24, no. 3, pp. 1–12, 2024.

[2] [S.-Y. Moon *et al.*, "Prediction of Type 2 Diabetes Using Logistic Regression in a Korean National Cohort," *Diabetes & Metabolism Journal*, vol. 45, no. 6, pp. 878–889, 2021.

[3] V. Vakil, S. Pachchigar, C. Chavda, and S. Soni, "Explainable Predictions of Different Machine Learning Algorithms Used to Predict Early Stage Diabetes," *arXiv preprint*, arXiv:2111.09939, 2021.

[4] S. S. Bhat *et al.*, "Prevalence and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandipora," *Health Science Reports*, vol. 5, no. 5, e723, 2022.

[5] S. G. Ugboaja *et al.*, "Advanced Diabetes Prediction Using Supervised Machine Learning Technique: Random Forest," *International Journal of Scientific Research and Engineering Development*, vol. 7, no. 1, pp. 314–320, 2024.

[6] Z. E. Huma, N. Tariq, and S. Zaidi, "Predictive Machine Learning Models for Early Diabetes Diagnosis: Enhancing Accuracy and Privacy with Federated Learning," *Journal of Computer-Based Instruction*, vol. 18, no. 2, pp. 45–53, 2024.

[7] M. A. Bülbül, "A Novel Hybrid Deep Learning Model for Early Stage Diabetes Risk Prediction," *Journal of Supercomputing*, vol. 80, no. 3, pp. 1412–1431, 2024.

[8] M. Jahangir *et al.*, "ECO-AMLP: A Decision Support System Using an Enhanced Class Outlier with Automatic Multilayer Perceptron for Diabetes Prediction," *arXiv preprint*, arXiv:1706.07679, 2017.

[9] M. Koumousis *et al.*, "Deep Learning Approach for Diabetes Prediction Using PIMA Indian Dataset," *Journal of Medical Systems*, vol. 44, no. 10, pp. 1–10, 2020.

[10] A. Ashiquzzaman *et al.*, "Reduction of Overfitting in Diabetes Prediction Using Deep Learning Neural Network," in *Proc. 20th International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2017, pp. 1–6.

[11] Anonymous, "Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes," *BioMed Research International*, vol. 2022, Article ID 101112, 2022.

[12] M. Owess *et al.*, "Prediction of Blood Sugar Level Using Supervised Machine Learning-Based Models," *Electronics*, vol. 13, no. 1, p. 92, 2024.

[13] Z. E. Huma *et al.*, "Federated XGBoost for Diabetes Diagnosis Using Distributed Health Records," *Journal of Computer-Based Instruction*, vol. 18, no. 2, pp. 60–67, 2024.

[14] F. Mohsen and Z. Shah, "Improving Early Prediction of Type 2 Diabetes Mellitus with ECG-DiaNet: A Multimodal Neural Network," *arXiv preprint*, arXiv:2504.05338, 2025.

[15] N. Tripathy, A. Samanta, and B. P. Mishra, "A Comparative Analysis of Diabetes Prediction Using Machine Learning and Deep Learning Algorithms in Healthcare," *International Journal of Information Technology and Computer Science*, vol. 16, no. 1, pp. 35–44, 2024.

NOTE: Customize as per your requirements.