# Meta-Analysing Diabetes Mellitus Discovery Using Artificial Intelligence Techniques

**Saravanan K.[1], Mohan V.[2]**

[1]Assistant Professor, Department of Computer Science and Engineering, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Samayapuram, Trichy, India.

[2]Professor, Department of Electronics and Communication Engineering, Saranathan College of Engineering, Trichy, India.

**E-mail:** [1]saravanank.set@dsuniversity.ac.in, [2] mohan-ece@saranathan.ac.in

## Abstract

In recent years, to provide more effective and timely treatment, accurate disease prediction for humans has remained highly challenging. Globally, diabetes is a multidisciplinary disease that threatens people's lives. It mainly targets the organs in the human body, such as the heart, kidneys, nerves, and eyes. By combining healthcare datasets with data fusion techniques and efficient machine learning models, a smart healthcare recommendation system can accurately forecast and suggest diabetes management. There have been recent proposals for machine learning models and methodologies to predict the onset of diabetes. However, these algorithms must remain current when dealing with the diversity of diabetes-related multi-feature datasets. Patients with uncontrolled diabetes are at increased risk of developing diabetes mellitus (DM), a disorder that can endanger many organs. Furthermore, this study delves into the interpretative capabilities of an ML model and the impact of its key components on prediction outcomes. Methods include gathering and organizing large datasets, including demographics, environmental variables, and past medical records. This work also examines the complex interactions with data and identifies patterns that indicate an epidemic of diabetes. The increasing rate of diabetes necessitates effective measures to prevent and manage potential diseases. This research proposes a novel framework for predicting patient diabetes using different machine learning algorithms that utilize

*DOI: https://doi.org/10.36548/jucct.2025.2.008*
*Received: 03.06.2025, received in revised form: 01.07.2025, accepted: 14.07.2025, published: 22.07.2025*

Electronic Health Record datasets for experimental purposes. The experimental work, carried out with various evaluation metrics, showed that it outperformed previous methods.

**Keywords:** Machine Learning, Decision Tree, Diabetes Mellitus, Electronic Health Records Datasets, Support Vector Machine.

## 1. Introduction

Significant advances in bioscience and, particularly, high-performance processors that consistently provide quick and affordable data creation enable machine learning in biology during the big data evolution. The main goals are to improve understanding of the increasing biological data and to provide better solutions to fundamental biological and medical concerns. The effectiveness and accuracy of these methods can be enhanced by utilizing the correct methodologies for developing recognition and data model generation. Treatment and prognosis of life-threatening diseases pose one of the biggest obstacles to scientific world. One example of such a disease is [1] diabetes mellitus (DM). Diabetes is a chronic disorder that can lead to high blood sugar levels and cause problems such as organ damage and failure.

The World Health Organization estimates that 7.3 million people in India are living with diabetes. Diabetic rates range from 10.7% to 14.2% in urban areas and 3.8 to 6.8% in rural areas. The national government estimates that 4.25 crore people in India have diabetes, with 25% of people living in the country. According to the NFHS-4, the number of people diagnosed with diabetes increased by 100% between 2007 and 2008. Among those aged 35 to 49, 3.5% of males and 3.5% of females have diabetes. In India, 9.7% of those aged 80 and older have diabetes, 13.1% of those aged 60–69 have it, and 13.2% of those aged 70–79 have it as well. The problem has worsened with age due to the mortality of patients. Approximately 36%, or one-third, of the remaining population has diabetes compared to those between the ages of 60 and 79.

The rising prevalence of diabetes among younger individuals poses serious health risks, including cardiovascular disease, stroke, renal disease, and blindness. There are different types of diabetes. Insulin generated by the liver and pancreas to work in type 1 diabetes. This hormone helps in the digestion of carbohydrates and lipids. Cells in type 2 diabetes are linked to the intestines, which prevents the body from processing insulin correctly. Over time, the body stops producing insulin, leading to the ineffectiveness and eventual failure of numerous bodily functions, ultimately resulting in death.

An increase in a female's blood sugar level, a symptom of type 3 diabetes, is linked to pregnancy [2,3]. There is no way to prepare for a type 1 rescue in advance. The most effective methods of preventing type 2 diabetes which accounts for 90% of cases, are a healthy diet, frequent exercise, and maintaining a healthy weight [4]. In 2014, out of 4.21 crore individuals who had medical examinations, 31 lakhs, or approximately 7.75%, were found to have diabetes. People still don't know much about diabetes. 40% of affected individuals cannot frequently care for themselves and cannot manage their sugar levels properly. Although they are at risk for retinal degeneration, half of them have never had an eye exam**.**

This research work gives important data on the relative performance of various supervised machine learning algorithms to help researchers choose the best algorithm for predicting disease outcomes. This study analyzes the effectiveness of several machine learning algorithms for illness risk prediction and reveals major patterns in forecasting diseases using machine learning. The field of data mining has recently shown interest in this topic. The research team reviewed Scopus and PubMed, two publicly accessible databases, in search of papers that used various supervised ML methods to forecast the onset of diseases. The success of these techniques stems from their capacity to derive models and patterns from data. Machine learning is particularly important in the period of big data the datasets can reach terabytes or petabytes in size. As a result, data-oriented research in biology has been greatly enhanced by the sheer volume of data.

Predicting and diagnosing diseases that endanger human health or impair quality of life is a crucial area of research in this hybrid field. Diseases like diabetes mellitus (DM) exemplify this. An important strategy to gain insights from the vast amounts of diabetes-related data is the application of data mining and machine learning techniques in DM research. Diabetes is one of the top research priorities in the medical due to its enormous societal impact, which, in turn, leads researchers to collect massive quantities of data. Consequently, DM diagnosis, treatment, and associated clinical management concerns revolve around data analysis and machine learning technologies. Therefore, this study aimed to examine diabetes research using machine learning and data mining techniques.

The rest of the work is categorized as follows: Section II discusses the related works on diabetes datasets used in this proposed study; Section III presents the proposed architectural framework; Section IV discusses different classification algorithms used in the HER dataset; Section V discusses the evaluation metrics used in classification algorithms; Section VI presents the experimental work; and the final section presents the conclusion.

## 2. Related Work

### 2.1 Structuring the Paper

First, the PIMA Indian Diabetes (PID) dataset, [14] conducted scientific experiments. The UCI machine learning repository offers 768 occurrences and eight variables. One of the most rapidly increasing long-term illnesses in 2014 was diabetes, which they intended to emphasize further, according to the World Health Organization (WHO). The success rates of the three classifiers utilized to determine if a person has diabetes were as follows: naive Bayes (77% accuracy), logistic regression (79% accuracy), and gradient boosting (86% accuracy).

Scientists [15] conducted experiments using a collection of data on diabetes obtained from the University of California, Irvine's database. In total, there were 520 instances with 16 different characteristics. They focused on premature diabetes prediction as their primary goal. It tackles problems and possibilities in the field of dynamic classification and prediction. It highlights the significance of multifaceted methods in addressing the complexity of nonlinear estimation and categorization and emphasizes the likelihood of studying multiple methodologies and approaches. Training results for various ML models showed that the RF classifier had the highest accuracy score of 98% on the relevant dataset. Logistic regression, support vector machines (SVMs), naive Bayes, decision trees, random forests, and multilayer perceptron followed with accuracy rates of 93%, 94%, and 98%, respectively. [16] They conducted experiments using a collection of data containing 520 individuals with diabetes and 17 characteristics that were retrieved from the UCI source.

They centered their efforts on diabetes detection in childhood. The SVM performs better in terms of both classification and recognition accuracy. With a rate of 93.27%, the naive Bayes classifier is the gold standard of classification algorithms. The best accuracy rate is 96.54%, achieved via SVM, while LightGBM is accurate only 88.46% of the time. Results like this prove that support vector machines are the preferred choice for forecasting who will develop diabetes.[17] They employed a 6-pronged approach to insulin recognition and analysis, including dataset processing methods, feature extraction, machine learning identification, and diagnosis and classification of DM to circumvent classification errors. Different controlled and unstructured methods, as well as grouping approaches, are compared. Improving the accuracy of diagnosis of various diabetic illnesses requires significant effort, and multiple databases present distinct obstacles.

Newly developed wrapper-based feature selection approaches optimized the multilayer perceptron (MLP) with the help of adaptive particle swarm optimization (APSO) and black wolf optimizer (BWO), thereby reducing the number of input characteristics required. In addition, the outcomes produced by this approach were contrasted with those of other conventional machine learning techniques, such as support vector machines (SVMs), decision trees (DTs), k-NNs, naïve Bayes classifiers (NBCs), random forest classifiers (RFCs), and logistic regression (LRs). The success rate achieved by LR was 95%. Regarding accuracy rates, k-NN came in at 96%, SVM at 95%, NBC at 93%, and both DT and RFC at 96%. The computational results of the proposed methods demonstrate that higher prediction accuracy can be achieved with fewer characteristics. Hopefully, this study will one day be useful to medical professionals in clinical practice.

Due to the dangerous nature of diabetes, early identification is consistently a challenge, according to Shafi et al. [18]. The authors of this work used classification methods from artificial intelligence to create a model that could detect the onset of diabetes early and address any issues. A paradigm that might reliably predict patients' chances of developing diabetes was developed by the authors of this research. This study examined and evaluated DT, SVM, and NBC, three machine learning classification algorithms, using a variety of metrics. Researchers could reduce time and achieve more accurate results by using the PID dataset obtained from the UCI repository. With an accuracy of 74%, the NBC method was deemed sufficient according to the trial data. SVM came in second with 63% precision, and DT came in third with 72% precision. Both the constructed infrastructure and the ML classifiers employed may find future utility in detecting and diagnosing other diseases. Researchers hoped to use the study and other ML approaches to better understand diabetes and planned to use it to categorize algorithms with insufficient information.

Diabetes is among the most serious illnesses globally in the modern era. People of all ages, from infants to older adults, are susceptible to contracting the condition. According to the International Diabetes Federation, in 2017, diabetes impacted 480 million people globally. However, studies indicate that the number of individuals expected to be affected is rising and could double by 2045. It is crucial to have diabetes predictions to prevent and manage the condition [5]. Reducing the number of people affected and eliminating the disease earlier than planned are possible outcomes of early diabetes prediction and prompt medication [6].

Once blood sugar levels are consistently high, a condition known as diabetes develops. The condition becomes life-threatening if not treated, necessitating insulin support throughout an individual's lifespan. Cardiovascular disease, liver problems, kidney failure, eye trouble, and other complications can develop as a result of the disease. Different forms of diabetes mellitus are defined with some specific limitations. Prediabetes, GDM, T1DM, and T2DM are the main forms. [7].

Impairment of glucose tolerance is another name for prediabetes. Glucose levels are elevated, although not as high as in type 2 diabetes. It will progress to type 2 diabetes mellitus if not addressed and managed in its early stages. Metabolic syndrome, characterized by low HDL levels and high blood pressure, might also develop as a result. Some tests that can detect prediabetes are the A1C, FPG, and OGTT [8]. One name for type 1 diabetes is juvenile diabetes (JDDM). Insulin deficiency occurs when the immune system destroys insulin-releasing cells, reducing the body's ability to produce insulin. Issues with the skin, heart, blood vessels, gums, nerves, pregnancy, retinopathy, and kidneys are among the complications that can arise from this condition, which often manifests during adolescence [9]. Autoantibodies in the blood and ketones in the urine are the diagnostic tools for type-1 diabetes.

Milder than type 1 diabetes, type 2 diabetes (T2DM) mostly affects the elderly. This condition develops when insulin production is reduced, or when the body develops antibodies that target insulin. [10]. It is also known as diabetes, that develops in adults. Common factors include obesity, a sedentary lifestyle, and compromised neurological and immunological systems, etc., are common factors. Eyes, nerves, kidneys, cardiovascular disease, and stroke are other complications. Testing for A1C, FPG, or RPG (Random Plasma Glucose) can help with the diagnosis. Glucose challenge and oral glucose tolerance tests are among the others [11].

Obstetric Diabetes Mellitus [24] is identified during the first several days of pregnancy. Issues including preterm delivery, high or low blood pressure, respiratory difficulties, complicated delivery weight, and diabetes later in life are among the problems. Both the glucose challenge and tolerance tests can detect diabetes. Diet, physical activity, insulin administration, and blood sugar monitoring are all part of diabetes management. For example, in most cases of diabetes during pregnancy, the likelihood of the condition developing after birth is low or non-existent. Nevertheless, it may recur in the affected person if it is not adequately addressed.

Gestational diabetes, which can occur during pregnancy, typically manifests itself in the second trimester. During gestational diabetes, the body either stops using insulin effectively or does not produce enough insulin to keep blood sugar levels stable. Overweight, a sedentary lifestyle, polycystic ovarian syndrome, genetics, and other risk factors are among the contributors. Hypoglycemia sensitivity testing and sugar stress testing are two types of examinations. Cesarean section, low blood sugar, preeclampsia, and postpartum type 2 diabetes are some of the risks. It typically goes away once the person delivers, but if untreated, it can lead to type 2 diabetes.

## 3. Principal and Methods Used in Machine Learning Algorithm

Forecasting diabetes outbreaks in urban areas is a crucial undertaking that requires employing advanced machine learning algorithms to discover outbreak patterns for timely notification to health organizations. This research work employs various classification algorithms, including Logistic Regression, Decision Trees, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Gradient Boosting Algorithm to enhance the accuracy and interpretability of diabetes outbreak predictions. The proposed model presents a comprehensive outline of a system designed for predicting diabetes outbreaks. This system encompasses multiple stages, beginning with data retrieval and culminating in model training, with the objective of offering businesses valuable insights into and foresight regarding diabetic prediction. The system examines architectural components. Central to this system is the procedure of extracting data from a database. The author obtains data from Kaggle, emphasizing the importance of effective data management for analytical purposes. The proposed architecture is shown in below figure 1.

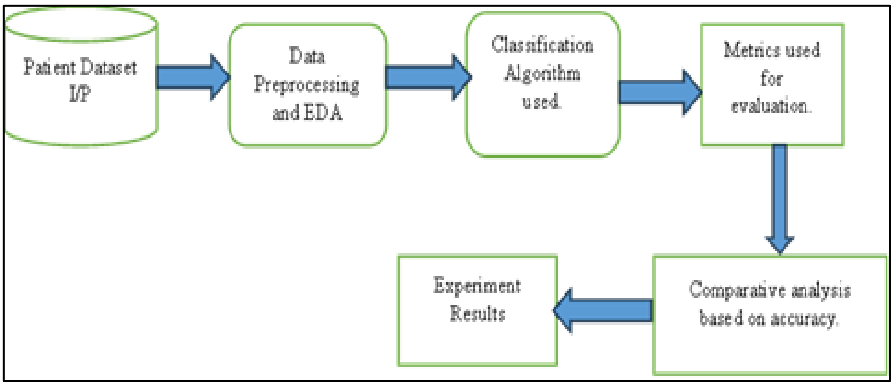
Patient Dataset I/P
Data Preprocessing and EDA
Classification Algorithm used.
Metrics used for evaluation.
Comparative analysis based on accuracy.
Experiment Results

**Figure 1.** Proposed Architecture

### 3.1 EHR Datasets

A large body of patient medical records kept digitally is known as an electronic health record (EHR) dataset. The information in these records includes a patient's diagnosis, prescriptions, treatment plans, dates of immunizations, allergies, images from radiology and laboratory tests, and much more. While the specifics inside an electronic health record (EHR) dataset may differ from one healthcare system to another, they overall cover a vast array of data pertaining to a patient's health and communications with medical professionals. The primary objective of these datasets is to identify and exploit the more prevalent elements in the incidence of disease in individuals, specifically the EHR dataset that contains attributes needed for diabetes prediction. This dataset comprises randomly collected information from an electronic health record database. It includes a total of 100000 rows, each representing an individual patient, with details recorded across 9 columns. The dataset encompasses various attributes, such as gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, glucose level, and diabetes.

### 3.2 Data Preprocessing

An essential part of getting EHR datasets ready for diabetes prediction or any other kind of predictive modeling is data pretreatment. Data must be clean, consistent, and appropriately preprocessed to train machine learning models. The main processes for preparing electronic health record (EHR) datasets for diabetes prediction involve handling missing data, verifying that the distribution of the dependent variable is suitable for classification, and ensuring that numerical parameters remain within the norm throughout training to avoid features with excessive scales overwhelming the process. Improve the model's forecasting abilities by adding new features. Data preparation refers to the steps taken to prepare the data for use in the computation process. The term refers to the steps used to transform unprocessed data into a more usable format [23].

### 3.3 Exploratory Data Analysis and Data Visualization

One of the most important functions of exploratory data analysis (EDA) is to help us recognize the features and patterns present in EHR datasets. Graphical and analytical exploration of the data is an integral part of EDA, an ongoing and iterative procedure that precedes explicit modeling. Data visualization is essential for obtaining information. This module highlights the significance of visualizing data and how it can reveal patterns, structures, and anomalies in diabetes behavior.

Visualizing the distribution of key features and identifying patterns: Learn to create visualizations that reveal the distribution of important features and patterns in the data. Interpreting the correlation matrix to discover feature relationships: The concept of a correlation matrix is introduced, and learners explore how it can uncover relationships between features.

This knowledge is crucial for feature selection and understanding feature importance in diabetes prediction. Figure 2 shows the chart representation of the frequency of hemoglobin levels to be listed directly below the names of the users. Multiple affiliations should be marked with superscript Arabic numbers, and they should start on a new line as shown in this document. In addition to the name of the affiliation, the system will ask users to provide the town and country in which it is situated, not including the entire postal address. E-mail addresses should start on a new line and should be grouped by affiliation.

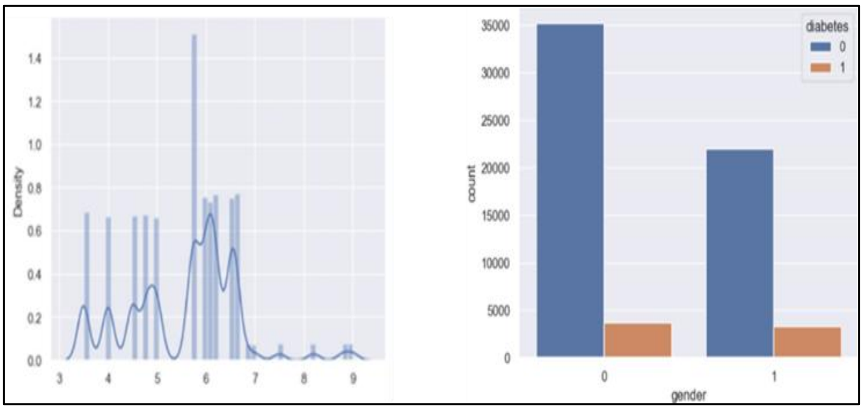

**Figure 2.** Frequency of Haemoglobin Levels and Cart Displaying Genders with Having Diabetes

## 4. Classification Algorithms

Predicting results from EHR datasets relies heavily on classification techniques. These algorithms classify patients' health states, risk factors, and additional applicable standards into different groups. Each problem's features, dataset properties, and forecast task's aims discuss the classification method. Hence, understanding the model's usability could be a crucial component in comprehending the reasons behind predictions, which is essential when making decisions in the healthcare industry.

### 4.1 Logistic Regression

It is a technique that is frequently used to model the probability of an event occurring based on one or more predictor variables. Logistic regression can be a useful technique for determining and comprehending the variables that influence the incidence of diabetes cases in a population when used to forecast a diabetes outbreak. The first step in using logistic regression for diabetes outbreak prediction is to collect pertinent data. Table 1 shows the classification report and accuracy of the logistic model.

**Table 1.** Classification Report and Accuracy Result of the Logistic Model

| | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|
| 0 | 0.96 | 0.99 | 0.97 | 17080 |
| 1 | 0.87 | 0.64 | 0.74 | 2172 |
| Accuracy | 0.92 | 0.89 | 0.95 | 19252 |
| Macro Avg | 0.91 | 0.81 | 0.86 | 19252 |
| Weighted Avg | 0.95 | 0.95 | 0.95 | 19252 |

Considerations such as sex, waist circumference, tobacco use, history, and other pertinent variables may be included in this data. Furthermore, information on the number of diabetes cases over time in a particular population is essential for both model validation and training. In figure 3 shows the ROC curve of the LR Tree and the confusion matrix representation.

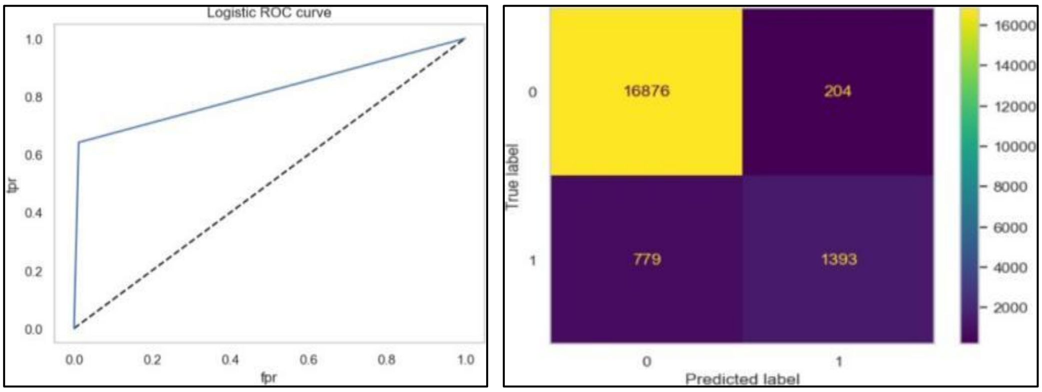

**Figure 3.** ROC Curve of Logistic Regression Tree and Confusion Matrix

The chosen variables are then used to train the logistic regression model. Using values for the selected predictors, the model calculates the likelihood that a person in the population has diabetes. The linear combination of predictors is often converted into a probability between 0 and 1 using the logistic function, also known as the sigmoid curve. To guarantee the model's dependability and predictive accuracy, its performance must be assessed after training. To evaluate how well the model generalizes to new, unseen data, a validation set of different data that was not used during training is used.

Logistic regression models are frequently assessed using metrics like the area under the receiver operating characteristic (ROC) curve, sensitivity, specificity, and accuracy. Public health interventions can be informed by the logistic regression model's predictions. To reduce the risk of a diabetes outbreak, targeted interventions, such as awareness campaigns, lifestyle modification programs, or early screening initiatives, can be implemented if specific factors are identified as significant predictors of diabetes.

Finally, the utilization of logistic regression offers a methodical and evidence-based technique for forecasting and comprehending the elements that lead to a diabetes pandemic. Researchers and public health officials can create proactive strategies to address the increasing problems associated with diabetes in a population by utilizing this statistical technique.

## 4.2 Support Vector Machine

Support Vector Machine uses the EHR dataset for gathering and compiling relevant information about predictor variables like age, BMI, family history, food habits, and physical activity, which constitutes the first stage. The individuals' diabetes status ought to be included in the dataset as well. Scaling the variables is crucial because SVM is sensitive to the feature scale. Standardization and normalization are two popular scaling strategies that ensure each feature contributes equally to the SVM model. The Radial Basis Function (RBF), among others, as well as quadratic and exponential kernel coefficients, can be used with SVM. The complexity of the relationships between variables and the type of data determine which kernel should be used. It might be necessary to evaluate various kernels to find one that works best for a particular dataset. By selecting support vectors and maximum scores, SVM creates the hyperplane from them. As a result, the concept of support vectors is born, giving rise to the SVM algorithm [19].

**Table 2.** Classifications Report and Accuracy Result of the SVM

| | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|
| 0 | 0.95 | 0.99 | 0.97 | 17080 |
| 1 | 0.92 | 0.61 | 0.73 | 2172 |
| Accuracy | 0.92 | 0.89 | 0.95 | 19252 |
| Macro Avg | 0.94 | 0.80 | 0.85 | 19252 |
| Weighted Avg | 0.95 | 0.95 | 0.95 | 19252 |

Table 2 shows the SVM report and accuracy for categorizing the test examples. SVM employs the Lagrangian formula described below. Figure 4 shows the ROC curve of the SVM tree and the confusion matrix.

$$S_{vm} = (U^{T}) = \sum_{i=1}^{1} £\, v_i a_i U_i U^{T} + B_o \qquad \text{-- (1)}$$

In the above equation the class labels of SVM are represented as $U_{i,}U^{T}$, $v_i, a_i$ are the test tuple, and $B_o$, *£,* are the numeric parameters [20]. Figure 4 represents the ROC curve of the SVM model.

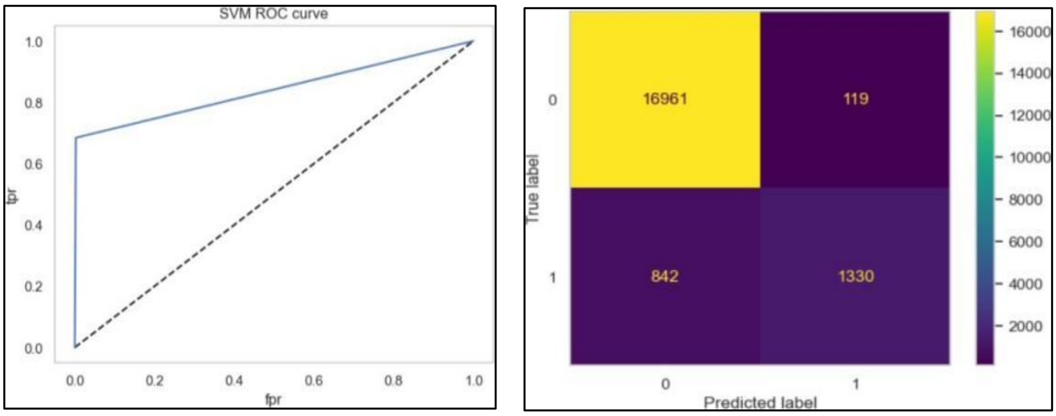

**Figure 4.** ROC Curve of SVM Tree and Confusion Matrix

### 4.3 Decision Trees

The effective machine learning method for anticipating and comprehending the occurrence of a diabetes outbreak is decision trees. Decision trees are helpful because they can handle both numerical and categorical data and can capture complex relationships between variables. Recursively dividing the data according to the predictor variables results in the construction of the decision tree. The below Table 3 shows the classification results and accuracy of the DT model. The algorithm chooses the variable that yields the best split at each node of the tree, maximizing the separation of people with and without diabetes. This process continues until a predefined stopping criterion, such as a smaller number of samples in each

**Table 3.** Classification Result and Accuracy of the DT Model

| | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|
| 0 | 0.96 | 0.97 | 0.97 | 17080 |
| 1 | 0.75 | 0.74 | 0.74 | 2172 |
| Accuracy | 0.92 | 0.89 | 0.94 | 19252 |
| Macro Avg | 0.86 | 0.85 | 0.86 | 19252 |
| Weighted Avg | 0.94 | 0.94 | 0.94 | 19252 |

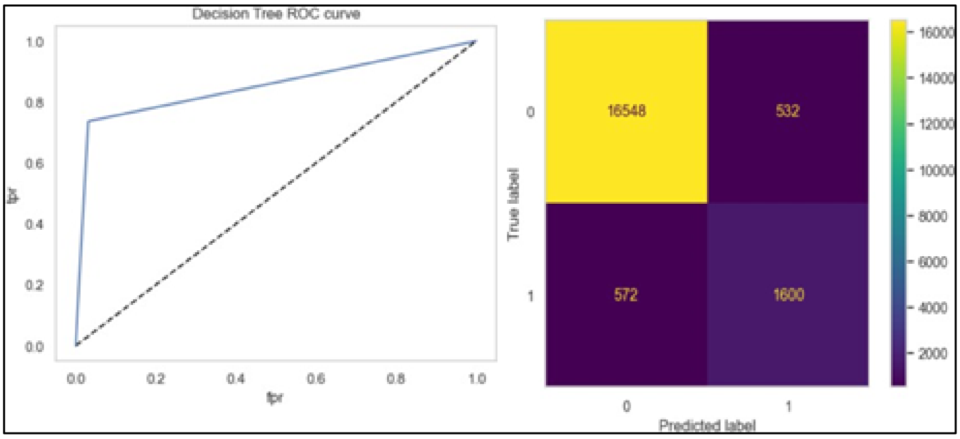

**Figure 5.** ROC Curve of Decision Tree and Confusion Matrix

The decision tree is generated once a leaf node or a predetermined tree depth is reached. The process for generating the decision tree is outlined as follows. Figure 5 illustrates the Receiver Operating Characteristic (RoC) curve of the decision tree, as well as the corresponding confusion matrix.

Entropy: We select a node ‘n’ and find the labels for the $C_L$ classes. The possible values of C are between 1 and $C_L$.

$$\text{Enpy(n)} = -\sum \; P(c|n)(c|n) \qquad \text{-- (2)}$$

## 4.4 Random Forest (RF)

With the goal of achieving a single outcome, the RF combines the results of multiple decision trees. In this case, the DTs are considered for both the base row and column sampling techniques.   when responding to the inputs, and the variation is minimized to enhance precision. One of the key bagging procedures is considered [21].

Random forest = Decision $_{\text{Tree}}$ (base beginner) + bagging (row selection with substitution) + feature bagging (column selection) + combination (mean/median, popular vote).

## 4.5 Gradient Boosting Algorithm

Gradient Boosting is an ensemble learning technique that builds a predictive model in the form of an ensemble of weak learners, typically decision trees. It is a powerful algorithm known for its ability to handle complex relationships in data and produce highly accurate predictions. Gradient Boosting creates an ensemble of decision trees in a stepwise manner, with each tree fixing the mistakes of the one before it. To increase prediction accuracy, the algorithm begins with a basic model, usually a shallow tree, and more trees are added as necessary.

**Table 4.** The Classification Result and Accuracy of the Gradient Boosting Model.

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.96 | 1.00 | 0.98 | 17080 |
| 1 | 0.97 | 0.68 | 0.80 | 2172 |
| Accuracy | 0.92 | 0.83 | 0.96 | 19252 |
| Macro Avg | 0.97 | 0.84 | 0.89 | 19252 |
| Weighted Avg | 0.96 | 0.96 | 0.96 | 19252 |

Table 4 shows the Classification result and Accuracy of the Gradient Boosting model.

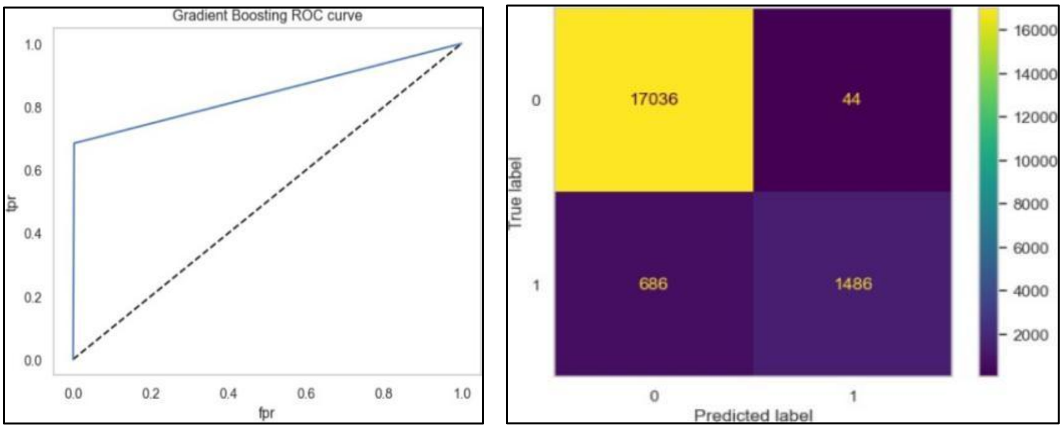

**Figure 6.** ROC Curve of Gradient Boosting and Confusion Matrix

Combining numerous weak learners into a model for prediction, usually in the shape of decision trees, constitutes a gradient-boosting classifier. A dataset's value count determines the optimal number of trees. It is often employed to reduce the simulation's bias error. The numerical values of the coefficients are obtained using a gradient-descent technique. Finding the coefficient's value requires determining the loss function that was utilized. The formula is $(x1 - x1')^2$, where x1 is the value initially measured and x1′ is the number the model finally projected. Consequently, the real target, Gain(A), is used instead of x1′ [22]. Figure 6 shows the gradient-boosting model representation.

$$\text{Gain}_{n+1}\,(A) = G_n\,(A) + £_n H1(a, CO_n) \qquad \text{-- (3)}$$

$$LF1 = (x1 - x1')^2, \qquad \text{-- (4)}$$

$$LF2 = (A\text{- Gain}_{n+1}(A))^2 \qquad \text{--(5)}$$

### 4.6 KNN Algorithm

An easy-to-understand and powerful artificial intelligence technique for categorization and tasks is K-Nearest Neighbours (KNN**).** The process begins with gathering and preparing a dataset that contains pertinent data on predictor variables like age, BMI, family history, food habits, and physical activity, just like other machine learning algorithms. The status of each person with diabetes should also be included in the dataset.

**Table 5.** Classification Result and Accuracy of the KNN model

| | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|
| 0 | 0.95 | 0.99 | 0.97 | 17080 |
| 1 | 0.93 | 0.60 | 0.73 | 2172 |
| Accuracy | 0.92 | 0.78 | 0.95 | 19252 |
| Macro Avg | 0.94 | 0.80 | 0.85 | 19252 |
| Weighted Avg | 0.95 | 0.95 | 0.95 | 19252 |

Table 5 presents the Classification result and Accuracy of the KNN model KNN uses a distance metric to calculate how similar two data points are to one another. Manhattan distance, the Euclidean distance, and other user-defined measures are among the most popular distance measures, albeit they vary by information type. Figure 7 shows KNN model ROC curve and confusion matrix.

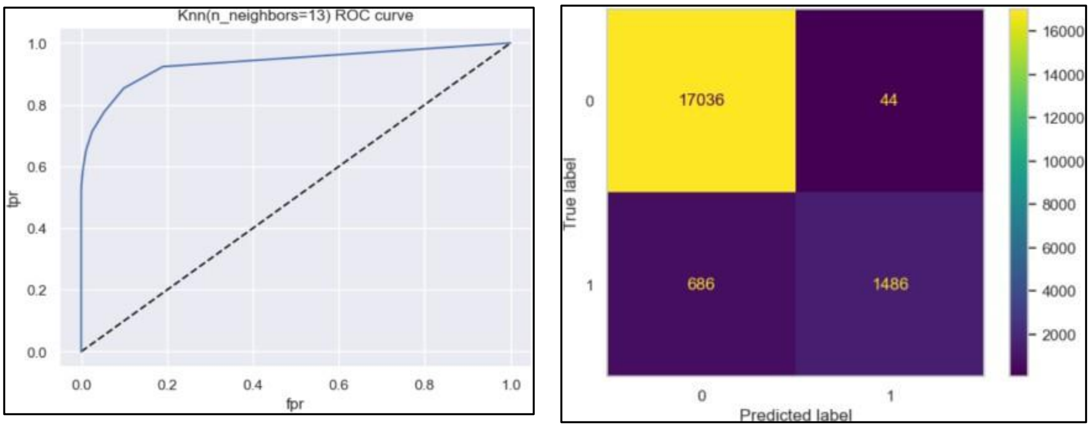

**Figure 7.** ROC Curve of KNN and Confusion Matrix

The algorithm's performance may be affected by the distance metric selected, and depending on the features of the dataset, it might need to be adjusted. The number of nearest neighbors that the algorithm considers when generating a prediction is represented by the "K" in KNN. Achieving a balance in the bias-variance tradeoff is essential when choosing a value for K. While larger K values can provide smoother decision boundaries, they may miss local patterns, whereas smaller K values may result in more flexible models that are more susceptible to noise.

## 5. Model Selection and Training

This proposed model provides an overview of the machine learning algorithms that will be employed for diabetes prediction. Each algorithm's characteristics, strengths, and weaknesses are detailed.

The Role of Logistic Regression as a Baseline Model: Logistic Regression is introduced as a fundamental classification algorithm. It enables users to understand the probability of diabetes based on linear combinations of input features. Its simplicity and interpretability make it an ideal starting point for this project.

Understanding Decision Trees to Capture Complex Patterns in Data: Decision Trees offer a non-linear approach to diabetes prediction. They recursively learn to split the dataset based on feature conditions, effectively capturing complex patterns and relationships in diabetic behavior. K-Nearest Neighbors (KNN) is utilized to forecast diabetes by sorting instances in the feature space according to the class that most closely matches the immediate neighbors of an instance. By considering similarities between data points, KNN can effectively distinguish between diabetic and non-diabetic cases.

Utilizing Support Vector Machines (SVM) for Defining Optimal Decision Boundaries: SVM is recognized as a powerful algorithm for establishing optimal decision boundaries. It seeks to identify the hyperplane that maximizes the margin between instances of diabetes and non-diabetes. Its versatility in handling high-dimensional data and non-linear relationships is particularly noteworthy.

Gradient Boosting for Diabetic Prediction: This method involves sequentially training weak models, typically decision trees, to rectify the errors of previous models. The algorithm prioritizes minimizing the difference between actual outcomes and predictions. By integrating these weak models and assigning greater weight to accurate predictions, a robust predictive model is developed for determining whether an individual is diabetic based on input features such as glucose levels, BMI, and hemoglobin levels.

## 6. Experimental Work

The proposed module provides an overview of the experimental work of predicting diabetes. K-Nearest Neighbors (KNN) in diabetes prediction involves classifying an example.

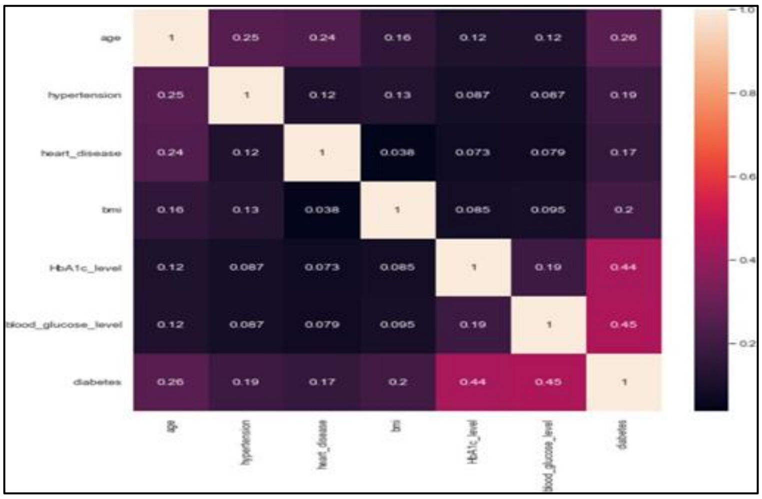

**Figure 8.** Correlation Matrix

Figure 8 presents the   result of the correlation matrix of the EHR data sets used in the evaluation of the models. The analysis of the correlation matrix reveals a notable positive correlation between "diabetes" and "blood_glucose_level." Consequently, it is feasible to include either one of them in the dataset, given their high correlation. Gradient Boosting for diabetic prediction involves sequentially training weak models (usually decision trees) to correct the errors of the previous ones. The algorithm focuses on minimizing the difference between actual outcomes and predictions. By combining these weak models and placing more weight on accurate predictions, a strong predictive model is created for determining whether an individual is diabetic based on input features like glucose levels, BMI, and hemoglobin levels.

**Table 6.** Comparison of All Classifier Performance Using Evaluation Metrics

| Method | Accuracy | Recall | Precision | F1-Score |
|---|---|---|---|---|
| K-NN | 95.33 | 97.8 | 95.54 | 97.22 |
| SVM | 95.67 | 98.20 | 95.65 | 97.12 |
| Logistic Regression | 94.81 | 98.21 | 96.23 | 97.41 |
| Decision Tree | 94.13 | 97.11 | 97.45 | 97.33 |
| Gradient Boosting | 96.20 | 97.12 | 96.66 | 98.45 |

Table 6 shows the comparison of all classifier performances using evaluation metrics. The machine learning algorithm is trained on the EHR dataset to provide diabetes mellitus predictions, and the methods with the best track records are selected and included in the Pima dataset. Table 1 shows the accuracy values of different classifiers and highlights the highest accuracy prediction achieved by the gradient boosting algorithm. Figure 9 presents a bar chart comparing the accuracy of different classification algorithms. The precision of the EHR information used to forecast diabetes mellitus differs among artificial intelligence classification algorithms. This outcome results from the data enhancement process for EHR datasets.

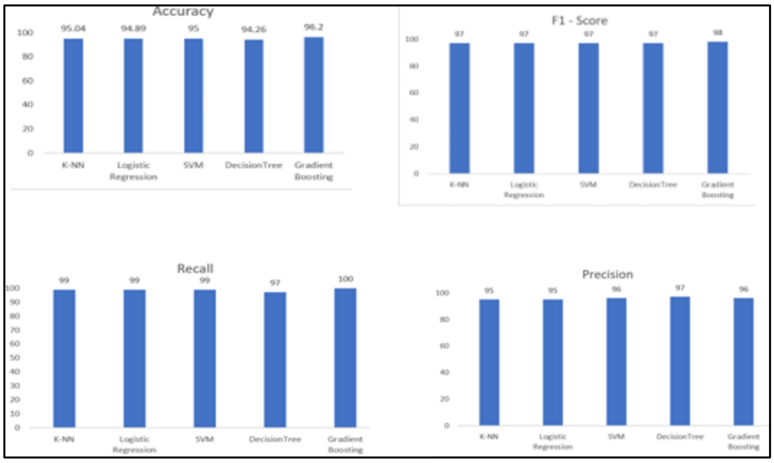

**Figure 9.** Classification Algorithm Accuracy Comparison

## 7. Conclusion

Finally, using artificial intelligence algorithms to discover meta-analysis of diabetes mellitus opens up exciting new possibilities for improving the basic knowledge and treatment of this complicated illness. Unlike conventional statistical methods, machine learning can analyze different datasets from various sources in search of patterns, correlations, and predictive models. Individualized treatment plans, early identification, and specific therapies based on patient profiles are all made possible by this comprehensive method.

Additionally, machine learning techniques provide doctors with advanced tools for data-driven decision-making, risk assessment, and therapy optimization. The goal of precision medicine in the fight against diabetes mellitus and improving patient outcomes is becoming closer to being achieved with each iteration and use of these approaches. These extensive analyses reveal that the GB approach differs from DT, RF, and SVM: according to previous studies, the GB approach outperforms other classification algorithms in terms of accuracy. The characteristics of the EHR dataset allowed the Gradient Boosting Algorithm to deliver outstanding results.

## References

[1] J. Chaki, S. T. Ganesh, S. K. Cidham, and S. A. Theertan, “Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review,” Journal of King Saud University. Computer and Information

Sciences/Mağalaẗ Ğam'aẗ Al-malīk Saud : Ùlm Al-ḥasib Wa Al-ma'lumat, vol. 34, no. 6, Jun. 2022, 3204–3225. doi: 10.1016/j.jksuci.2020.06.013.

[2] J. Spranger et al., “Adiponectin and protection against type 2 diabetes mellitus,” Lancet, vol. 361, no. 9353, Jan. 2003, 226–228. doi: 10.1016/s0140-6736(03)12255-6.

[3] B. B. Lowell and G. I. Shulman, “Mitochondrial dysfunction and type 2 diabetes,” Science, vol. 307, no. 5708, Jan. 2005, 384–387. doi: 10.1126/science.1104343.

[4] G. Klöppel, M. Löhr, K. Habich, M. Oberholzer, and P. U. Heitz, “Islet Pathology and the Pathogenesis of Type 1 and Type 2 Diabetes mellitus Revisited,” Pathology and Immunopathology Research, vol. 4, no. 2, Jan. 1985, 110–125.doi: 10.1159/000156969.

[5] Dr. D. Goutam et al., “A NEW APPROACH FOR OVARIAN CANCER MANAGEMENT,” YMER, vol. 22, no. 1, Jan. 2023, 928–953. doi: 10.37896/YMER22.01/73.

[6] A. H. Syed and T. Khan, “Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study,” IEEE Access, vol. 8, Jan. 2020, 199539–199561. doi: 10.1109/access.2020.3035026.

[7] Global Atlas of Artificial Intelligence Regulation. 2022 Edition: Oriental Vector. 2022. doi: 10.24866/7444-5326-8.

[8] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. García-García, “Machine learning and deep learning predictive models for type 2 diabetes: a systematic review,” Diabetology & Metabolic Syndrome, vol. 13, no. 1, Dec. 2021, doi: 10.1186/s13098-021-00767-9.

[9] N. G. Ramadhan, A. Adiwijaya, and A. Romadhony, “Preprocessing Handling to Enhance Detection of Type 2 Diabetes Mellitus based on Random Forest,” International Journal of Advanced Computer Science and Applications, vol. 12, no. 7, Jan. 2021, doi: 10.14569/ijacsa.2021.0120726.

[10] S. I. Ayon and Md. M. Islam, “Diabetes Prediction: a deep learning approach,” International Journal of Information Engineering and Electronic Business, vol. 11, no. 2, Mar. 2019, 21–27. doi: 10.5815/ijieeb.2019.02.03.

[11] P. Uppamma and S. Bhattacharya, “Deep Learning and Medical Image processing Techniques for Diabetic Retinopathy: A survey of applications, challenges, and future trends,” Journal of Healthcare Engineering, vol. 2023, Feb. 2023, 1–18. doi: 10.1155/2023/2728719.

[12] Y. Deng et al., “Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients,” Npj Digital Medicine, vol. 4, no. 1, Jul. 2021, doi: 10.1038/s41746-021-00480-x.

[13] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, “Applications of machine learning to machine fault diagnosis: A review and roadmap,” Mechanical Systems and Signal Processing, vol. 138, Apr. 2020, 106587. doi: 10.1016/j.ymssp.2019.106587.

[14] M. Abu-Farha, J. A. Abubaker, and J. Tuomilehto, Diabetes in the Middle East. Frontiers Media SA, 2021.

[15] D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu, Nonlinear estimation, and classification. Springer Science & Business Media, 2013.

[16] J. Xue, F. Min, and F. Ma, “Research on Diabetes Prediction Method based on Machine learning,” Journal of Physics: Conference Series, vol. 1684, Nov. 2020, 012062. doi: 10.1088/1742-6596/1684/1/012062.

[17] J. Chaki, S. T. Ganesh, S. K. Cidham, and S. Ananda (eertan,“Machine learning and artificial intelligence based diabetesmellitus detection and self-management: a systematic review,”Journal of King Saud University - Computer and Information Sciences, vol. 34, 2020, 1319–1578.

[18] T. R. Gadekallu et al., “Early detection of diabetic retinopathy using PCA-Firefly based deep learning model,” Electronics, vol. 9, no. 2, Feb. 2020, 274. doi: 10.3390/electronics9020274.

[19] J. M. Machado et al., Distributed Computing and Artificial Intelligence, Special Sessions, 19th International Conference. Springer Nature, 2023.

[20] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavão, and K. P. Bennett, “Generation and evaluation of privacy preserving synthetic health data,” Neurocomputing, vol. 416, Nov. 2020, 244–255. doi: 10.1016/j.neucom.2019.12.136.

[21] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: review, opportunities and challenges,” Briefings in Bioinformatics, vol. 19, no. 6, May 2017, 1236–1246. doi: 10.1093/bib/bbx044.

[22] S. Habibi, M. Ahmadi, and S. Alizadeh, “Type 2 diabetes mellitus screening and risk factors using Decision Tree: Results of Data Mining,” Global Journal of Health Science, vol. 7, no. 5, Mar. 2015, doi: 10.5539/gjhs.v7n5p304.

[23] S. S. Kshatri, K. Thakur, M. H. M. Khan, D. Singh, and G. R. Sinha, Computational intelligence and applications for pandemics and healthcare. IGI Global, 2022.

[24] S. Habibi, M. Ahmadi, and S. Alizadeh, “Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining,” Global Journal of Health Science, vol. 7, no. 5, 2015, 304–310.