DOI: 10.52756/ijeim.2025.v01.i01.006 Int. J. Eng. Infor. Manag. 1(1): 73-92 (2025)

International Journal of Engineering and Information Management
Journal homepage: www.ijeim.in

A Machine Learning Approach for Multiclass Classification of Diabetes

Soumen Ghosh[1*], **Souvik Halder**[1], **Prasun Maity**[1], **Shiladitya Munshi**[2]

[1]Department of Computer Science & Engineering (Data Science), Haldia Institute of Technology, West Bengal, India
[2]Department of Computer Science & Engineering, Techno India University, Tripura, Agartala, India

***Corresponding Author:** soumenghoshcse2812@gmail.com

Article Information

Type of Article: Original
Received: Oct 7, 2024
Accepted: Dec 30, 2024
Published: Jan 15, 2025

Keywords:
Diabetes
Machine Learning
Logistic Regression
SVM
Decision Tree
Random Forest
Gradient Boosting

Cite this article:
Soumen Ghosh, Souvik Halder, Prasun Maity and Shiladitya Munshi. (2025). A Machine Learning Approach for Multiclass Classification of Diabetes. *International Journal of Engineering and Information Management*, 1(1), 73-92.
DOI: 10.52756/ijeim.2025.v01.i01.006

Abstract

Diabetes is an increasing global health issue, with millions at risk due to factors like lifestyle, genetics, and other health conditions. Early diagnosis is essential for timely treatment, avoiding complications, and easing the strain on healthcare systems. The disease's complexity, with its different stages, requires advanced models that can distinguish between diabetic, non-diabetic, and pre-diabetic individuals. This study aimed to develop a precise multiclass classification model to predict a patient's diabetes status based on various health indicators. In addition to standard factors like blood sugar level, BMI, cholesterol, and age, external risk factors have also been considered for better accuracy. In the current study, the target variable categorizes patients as Diabetic, Non-Diabetic, or Pre-Diabetic. The current work applies Logistic Regression, SVM, Decision Tree, Random Forest, and Gradient Boosting models to address the classification challenge. After training and testing the models, Random Forest has been identified to deliver the highest accuracy at 98%, outperforming the others. These findings highlight the power of machine learning in effectively classifying patients based on diabetes status.

1 Introduction

The identification of diabetes is critically important for comprehensive healthcare management and individual well-being. Early detection allows patients to implement proactive lifestyle modifications and medical interventions that can significantly slow disease progression, potentially preventing or delaying severe complications like cardiovascular disease, kidney damage, and nerve disorders. Accurate identification enables healthcare professionals to develop personalized treatment strategies, prescribing appropriate medications, designing tailored diet and exercise plans, and monitoring specific risk factors unique to each diabetes type. Undiagnosed or poorly managed diabetes can lead to life-threatening health consequences, including kidney failure, vision problems, increased stroke risk, and impaired wound healing. Moreover, timely diabetes identification has broader implications for healthcare systems, reducing long-term medical costs, decreasing hospitalizations, and improving overall public health management. Beyond medical metrics, proper diabetes identification empowers patients with a deeper understanding of their health condition, enhancing their self-management capabilities, psychological preparedness, and potential for maintaining a high quality of life. By recognizing diabetes early and accurately, individuals can transform what could be a debilitating chronic condition into a manageable aspect of personal health, ensuring better long-term outcomes and preventing potentially severe medical complications. In 2023, a study conducted by the Indian Council of Medical Research (ICMR) and

reported in Anjana et al. (2023) mentioned that over 10.1 crore individuals in India are affected by diabetes. The increasing prevalence of diabetes has highlighted the urgent need for early detection and effective management strategies. Machine learning models have been identified as promising to enhance early diagnosis through accurate, non-invasive predictions, making them a crucial tool in modern healthcare.

The research gap addressed by the proposed study lies in the transition from traditional binary classification models to a multiclass classification approach for diabetes prediction. Typically, binary models classify patients as either diabetic or non-diabetic. However, an additional category, "Predict-Diabetic," has been introduced in the proposed multiclass model. This category aids in identifying individuals at high risk of developing diabetes before its full manifestation. By incorporating this advanced classification framework, a more refined understanding of diabetes risk is provided, enabling timely interventions and personalized care.

A unique combination of features has been employed in the dataset, distinguishing this study from prior research. While common health indicators such as Body Mass Index (BMI), age, and blood sugar levels are included, more comprehensive metrics such as creatinine ratio (Cr), cholesterol levels (total, LDL, VLDL, triglycerides, HDL), and fasting lipid profile have also been considered. This extensive set of features enhances the model's capability to capture a broader range of health patterns, leading to more accurate and reliable predictions.

Few machine learning algorithms like Logistic Regression, SVM, Decision Tree, Random Forest and Gradiant Boosting—have been applied for the implementation of the model. Each model has been rigorously evaluated based on key performance metrics, including accuracy and recall rates. The models demonstrated significant performance, with average accuracy ranging from 89% to 97%. The highest accuracy, 98%, was achieved by the Random Forest model, which also showed exceptionally high recall rates, indicating its effectiveness in correctly identifying diabetic cases. These results underscore the robustness of the model and its potential for real-world application.

While the results achieved with machine learning models in this study are impressive, it is worth noting that similar levels of accuracy have been achieved in other studies using more complex deep learning models. However, by employing traditional machine learning algorithms, comparable performance has been obtained, along with additional advantages of computational efficiency and interpretability. This makes the models particularly suitable for deployment in practical healthcare environments.

Furthermore, expanding the model to include genetic and lifestyle factors could provide a more comprehensive risk assessment. By integrating diverse data sources, including genetic predispositions and behavioral patterns, even more precise prediction models could be developed. Collaborative efforts between healthcare providers and researchers could lead to the creation of robust, multi-faceted models tailored to individual patient needs.

The current research addresses a critical gap in diabetes prediction while demonstrating the efficacy of machine learning models in achieving high accuracy and recall rates. With continued advancements and future enhancements, the model has the potential to significantly impact early diabetes detection and management, ultimately contributing to improved patient outcomes and more efficient healthcare delivery.

2 Literature Review

Diabetes is a prevalent and chronic condition that poses significant health risks globally. Early prediction and classification of diabetes are crucial for timely medical intervention. Machine learning algorithms have been widely adopted in recent years for the classification and prediction of diabetes. This section provides a survey of relevant studies that have employed machine learning methods for diabetes prediction, highlighting various models, techniques, and their respective performance.

2.1 Machine Learning Approaches for Diabetes Prediction

Mujumdar & Vaidehi (2019) introduced the application of machine learning algorithms for diabetes prediction, specifically exploring the effectiveness of classification models in predicting diabetes outcomes. Their study demonstrated the potential of algorithms like Decision Trees, Support Vector Machines (SVM), and Random Forests in predicting diabetes based on various medical parameters. Butt et al. (2021) further explored the use of machine learning models in healthcare, focusing on diabetes classification and prediction. They combined multiple classifiers, such as Random Forest and Logistic Regression, to improve the accuracy of diabetes prediction models. Their findings highlighted the importance of using multiple classifiers for robust healthcare applications. Qiao et al. (2020) focused on a specialized aspect of diabetes prediction by using deep learning algorithms to detect diabetic retinopathy, a complication of diabetes. Their model, which uses deep learning techniques on fundus images, not only detects diabetes but also predicts the risk of developing complications like retinopathy, providing an additional layer of predictive capability. Liang et al. (2021) proposed a radiomics-based approach for predicting diabetic foot conditions using fundus images. This study is particularly relevant as it extends diabetes prediction to complications associated with the disease, offering insights into the potential of image-based diagnostics in diabetes prediction.

2.2 Ensemble Techniques and Hybrid Models

Several studies have employed ensemble methods to enhance the performance of machine learning models for diabetes prediction. Hasan et al. (2020) utilized an ensemble approach by combining multiple classifiers to improve prediction accuracy. Their findings demonstrated that ensemble models significantly outperform individual models, providing more accurate predictions and better generalization. In a similar vein, Ayon & Islam (2019) used deep learning methods for diabetes prediction, demonstrating the advantages of leveraging complex neural networks for improving prediction accuracy. This approach, while computationally more intensive, offers high accuracy and robustness in predicting diabetes outcomes. Khanam & Foo (2021) compared several machine learning algorithms for diabetes prediction, including Decision Trees, SVM, and K-Nearest Neighbors (KNN). They concluded that while Decision Trees and SVM showed promising results, Random Forests outperformed them, providing a more robust and accurate classification model.

2.3 Early Prediction and Risk Assessment

Early prediction and risk assessment of diabetes are crucial for prevention and management. Alam et al. (2019) focused on developing models for the early prediction of diabetes. Their work integrated machine learning models with clinical data to predict diabetes onset in its early stages, thus aiding in preventative healthcare measures. Jayanthi et al. (2017) surveyed clinical prediction models for diabetes, discussing various machine learning models that have been employed to predict diabetes based on clinical parameters. They noted that while many models show promise, there is a need for further refinement in model interpretability and clinical applicability. Bukhari et al. (2021) proposed an improved Artificial Neural Network (ANN) model for diabetes prediction. They demonstrated that a well-designed neural network model could offer improved prediction accuracy by capturing complex patterns in diabetes-related data, a key advantage for clinical decision support.

2.4 Comparison of Classifiers and Model Selection

Nai-Arun & Moungmai (2015) compared classifiers like SVM, Naive Bayes, and K-Nearest Neighbors for predicting diabetes, finding SVM to be the most suitable for accurate classification. Xue et al. (2020) proposed a hybrid machine learning approach that combined SVM with artificial neural networks for diabetes prediction. Their study showed that hybrid models could leverage the strengths of multiple algorithms to improve predictive accuracy and robustness, particularly in complex and high-dimensional datasets. Ahmed et al. (2022) introduced a fused machine learning approach for diabetes prediction, where they combined different machine learning techniques to create a more powerful model. Their

approach enhanced prediction accuracy, showcasing the advantages of hybridization in machine learning models for healthcare applications.

2.5 Importance of Data Features in Diabetes Prediction

The selection of relevant features is essential for building effective prediction models. Studies by Carstensen et al. (2020) and Sinha & Lipton (2021) have demonstrated that age and glucose levels are key indicators for predicting diabetes onset. Fox & Flegal (2023) further emphasized the importance of BMI as a predictor for type 2 diabetes, with high BMI being a well-established risk factor. Some studies highlighted the role of creatinine levels in assessing kidney function in diabetic patients. Elevated creatinine levels are commonly associated with diabetic kidney disease, making them an important feature for prediction models targeting diabetic complications. Similarly, elevated urea levels, are another important indicator of renal dysfunction in diabetes, which can aid in both early detection and risk assessment.

The literature demonstrates a growing volume of work on using machine learning algorithms for diabetes prediction. Various approaches, from traditional machine learning algorithms like Random Forests and SVM to deep learning techniques, have been explored with promising results. Ensemble methods and hybrid models have shown great potential in improving prediction accuracy and generalization. Furthermore, the inclusion of key physiological features such as age, glucose levels, BMI, and creatinine ratio enhances the predictive capability of these models.

As the field continues to evolve, future research may focus on integrating more diverse data sources, such as genetic information and lifestyle factors, to develop more personalized and accurate prediction models. Moreover, there is significant potential for creating models that not only predict diabetes onset but also provide early detection of complications, improving patient outcomes through timely intervention.

3 Research Methodology

The current section discusses the problem formulation and the solution requirements to offer a generic guideline for machine learning model design principles for multiclass classification of diabetes data. On the basis of these guidelines or objectives, subsequent subsections report the justifications for selecting diabetes related parameters and machine learning models for diabetes data classification.. However, for the completeness of the paper, the current section mentions preliminary- level background studies on the different machine learning models used and diabetes data parameters.

3.1 Problem Formulation

The objective of this study is to develop a multiclass classification model that can accurately predict the diabetes status of patients. The target variable is the diabetes class, which categorizes patients as Diabetic, Non-Diabetic, or Predict-Diabetic. The prediction or classification is based on several health indicators, including standard factors such as sugar levels, BMI, cholesterol levels, and age, as well as external factors like the creatinine ratio and fasting lipid profile. All the variables including target classes and data parameters have been expressed with some specified notations as described in Table 1. These notations have been extensively used in next discussion which officially states the problem.

3.1.1 Problem Statement

Given m number of labeled instances having, personal, physiological and target class related information, to train the the chosen model(s) and an unseen test set of n instances of raw physiological information (augmented with personal information); design a multi-class classifier to detect Existence or Non-existence or Pre-Existence of diabetes.

Table 1: Notations used in the study.

Item	Notation	Remark
Patient_id	pat_id	Personal Information
Age	ag	Personal Information
Gender	gndr	Personal Information, values can be M/F/T
Sugar Level	sl	Physiological Information
Creatinine Ratio	cr	Physiological Information
Body Mass Index	bmi	Physiological Information
Urea	ur	Physiological Information
Cholesterol	ch	Physiological Information
Fasting Lipid Profile	flp	Physiological Information
Glycated Haemoglobin Test Vslue	hba1c	Physiological Information
Target Diabetic class	C_0	Represents Diabetic Class
Target Predict Diabetic class	C_1	Represents Predict-Diabetic Class
Target Non Diabetic class	C_2	Represents Target Non Diabetic Class

3.2 Justification of chosen physiological features

The early prediction of diabetes relies on multiple physiological parameters that indicate the risk and severity of the disease. The following are the critical parameters chosen for this purpose, each justified by evidence from existing research:

- **Age:** Age is a significant factor in diabetes risk, with the incidence of type 2 diabetes increasing with age. This relationship is attributed to metabolic changes and increased insulin resistance as individuals grow older Carstensen et al. (2020).
- **Gender:** Gender differences are known to affect diabetes risk, with women and men experiencing distinct risk profiles. Hormonal differences and lifestyle factors contribute to these variances, making gender a necessary consideration in predictive models.
- **Sugar Level:** Blood glucose levels are directly indicative of diabetes, as hyperglycemia is a defining characteristic of the disease. Monitoring sugar levels provides immediate insight into the body's ability to regulate glucose Sinha & Lipton (2021).
- **Creatinine Ratio:** The creatinine ratio is associated with kidney function, which is often impaired in diabetes due to hyperglycemia. Diabetes-induced kidney damage results in elevated creatinine levels, making this ratio a crucial indicator.
- **Body Mass Index (BMI):** High BMI is linked to increased risk for type 2 diabetes, as excess body weight contributes to insulin resistance. Studies demonstrate that obesity and elevated BMI are consistent risk factors Fox & Flegal (2023).
- **Urea:** Elevated urea levels are indicative of renal dysfunction, a common complication of diabetes. Monitoring urea can help detect early signs of kidney involvement in diabetic patients.
- **Cholesterol:** Dyslipidemia, characterized by abnormal cholesterol levels, is prevalent in diabetic patients. High cholesterol levels are associated with cardiovascular complications in diabetes, making this parameter essential for comprehensive risk assessment Gerstein & Ong (2024).
- **Fasting Lipid Profile:** The fasting lipid profile provides a detailed view of lipid abnormalities, including triglycerides and HDL/LDL cholesterol levels, which are often altered in diabetic patients. This profile is critical for understanding lipid metabolism disturbances related to diabetes Fruchart et al. (2021).

- **Glycated Hemoglobin (HbA1c) Test:** The HbA1c test measures average blood glucose levels over the past 2-3 months, providing a long-term view of glucose control. It is a standard diagnostic tool for diabetes and a reliable predictor of disease onset Wilson & Porter (2022).

3.3 Choice of Learning Models

Diabetes prediction is a critical task in healthcare, aiming to identify individuals at risk of developing diabetes mellitus, particularly type 2 diabetes. Accurate prediction models enable early intervention, reducing the prevalence of complications associated with the disease. This paper justifies the selection of specific machine learning models like Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, and Gradient Boosting—for diabetes prediction. Additionally, it discusses the theoretical limitations of other popular learning models in this context. The reasons behind the selection of learning models are given below.

(i) *Logistic Regression:* Logistic Regression is a fundamental classification algorithm widely used for binary outcomes, making it suitable for predicting the presence or absence of diabetes. The model provides clear insights into the relationship between input features and the probability of diabetes, which is valuable in clinical settings Smith & Johnson (2021). It requires relatively low computational resources, allowing for quick training and prediction. This model often serves as a baseline model against which more complex models are compared Davis & Lee (2019).

(ii) *Support Vector Machines (SVM):* SVMs are powerful classifiers known for their effectiveness in high-dimensional spaces. SVMs aim to maximize the margin between classes, enhancing generalization performance. This robust Huang & Li (2021) model has an ability to handle non-linear relationships by applying different kernel functions, making them versatile for complex diabetes data Vapnik (2013).

(iii) *Decision Trees:* Decision Trees offer a non-parametric approach to classification, suitable for diabetes prediction due to its clear and understandable model structure, facilitating decision-making in clinical practice Quinlan (2014). It is capable of managing both numerical and categorical features without the need for extensive preprocessing Hastie et al. (2009). Moreover, Decision Trees have the ability to model complex interactions between features Breiman (2001).

(iv) *Random Forest:* Random Forests, an ensemble of Decision Trees, enhance prediction accuracy and robustness by averaging multiple trees, thereby mitigating the risk of overfitting common in single Decision Trees Efron (2001). They provide measures of feature importance, aiding in the identification of significant predictors for diabetes Liaw & Wiener (2002). Randon Forest is capable of handling large datasets with higher dimensionality and complex feature interactions Breiman (2001).

(v) *Gradient Boosting:* Gradient Boosting machines, including algorithms like XGBoost and LightGBM, are highly effective for predictive tasks. They consistently achieve superior accuracy by sequentially correcting the errors of previous models Chen & Guestrin (2016). Additionally, it has the ability to optimize different loss functions and incorporate regularization techniques to prevent overfitting He & Zhang (2023). Gradient Boosting models are known for their effective in managing missing values within the dataset Shi & Zhang (2021), which is pretty common in healthcare sector.

Limitations of Other Popular Learning Models: While numerous machine learning models exist, some are less suited for diabetes prediction due to specific limitations:

(i) *Neural Networks:* Neural Networks typically require large datasets to perform well, which may not always be available in medical contexts LeCun et al. (2015). Often considered "black boxes," Neural Network makes it difficult to interpret the relationship between features and predictions which is a crucial aspect in healthcare Ribeiro et al. (2016). Training deep neural networks demands significant computational power and time, which limits it scalability, a mandatory condition for wide adoption in healthcare. Goodfellow et al. (2016).

(ii) *K-Nearest Neighbors (KNN):* KNN struggles with moderate to large datasets due to high computational and memory requirements Cover & Hart (1967). It also requires careful normalization of features, which can be cumbersome with mixed data types Cover & Hart (1967). Performance of KNN can degrade abruptly with noisy or irrelevant features Guyon & Gunn (2003), which is pretty common assumption in mass scale adoption of learning models in healthcare domain.

(iii) *Naive Bayes:* The strong assumption that features are independent often does not hold in medical data, leading to suboptimal performance Rennie (1997) in case of Naive Bayes. Moreover, it is considered as less capable of capturing complex relationships between features compared to other models like Random Forest or Gradient Boosting Rao & Narasimhan (2003).

Logistic Regression, SVM, Decision Trees, Random Forest, and Gradient Boosting are well-suited for diabetes prediction due to their balance of interpretability, predictive performance, and ability to handle complex data structures. While other models like Neural Networks, K-Nearest Neighbors, and Naive Bayes offer their own advantages, they present significant challenges in the context of diabetes prediction, such as interpretability issues, scalability problems, and restrictive assumptions. Therefore, the selected models provide an optimal combination of performance and practicality for effectively predicting diabetes.

For, the completeness of this study, following subsesctions provide fundamental accounts of the mathematical concept and formulation of the chosen learning model.

3.4 Logistic Regression

Model Overview

In Logistic Regression, we model the probability $P(Y = 1|X)$ that the dependent variable Y (for example, presence or absence of diabetes) equals 1, given the independent variable(s) $X = (X_1, X_2, \ldots, X_n)$. Logistic Regression uses the *logistic (sigmoid) function* to transform the output of a linear equation into a probability between 0 and 1.

Linear Combination of Inputs

First, we create a linear combination of the input features as described in Equation 1

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n = \beta^T X \tag{1}$$

where:

- β_0 is the intercept term (bias),
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for each feature $X_1, X_2, \ldots, X_n$,
- $\beta = [\beta_0, \beta_1, \ldots, \beta_n]$ and $X = [1, X_1, X_2, \ldots, X_n]$ are the coefficient and feature vectors, respectively.

Sigmoid (Logistic) Function

To ensure that the output is a probability (i.e., between 0 and 1), we apply the sigmoid function to z as expressed in Equation 2

$$P(Y = 1|X) = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

where $\sigma(z)$ denotes the sigmoid function.

Log-Odds Interpretation

In Logistic Regression, the *log-odds* or *logit* function is linear in the parameters as shown in Equation 3:

$$\text{logit}(P(Y=1|X)) = \ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta^T X \tag{3}$$

This equation states that the logarithm of the odds of the positive class is a linear function of the input features.

Cost Function (Log-Loss)

To estimate the parameters β, we use Maximum Likelihood Estimation (MLE), which aims to maximize the probability of observing the true labels in the training data. In practice, this is equivalent to minimizing the *logistic loss* or *binary cross-entropy loss* function as mentioned in Equation 4:

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\hat{y}^{(i)} + (1-y^{(i)})\log(1-\hat{y}^{(i)})\right] \tag{4}$$

where:

- $y^{(i)}$ is the actual label (0 or 1) of the i-th sample,
- $\hat{y}^{(i)}$ is the predicted probability for the positive class, given by $\sigma(\beta^T X^{(i)})$,
- m is the total number of training samples.

Gradient Descent for Optimization

The logistic loss function $J(\beta)$ is minimized using gradient descent or one of its variants, like stochastic gradient descent (SGD). The gradient of $J(\beta)$ with respect to the parameters β is shown in following equation Equation 5 :

$$\frac{\partial J}{\partial \beta_j} = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)X_j^{(i)} \tag{5}$$

where $X_j^{(i)}$ is the j-th feature value of the i-th sample.

This gradient is used to update the parameters in the direction that reduces the cost as per Equation 6 :

$$\beta_j := \beta_j - \alpha\frac{\partial J}{\partial \beta_j} \tag{6}$$

where α is the learning rate.

Decision Boundary

For a binary classification problem, the decision boundary is the threshold at which we decide the output class. In Logistic Regression, we typically use a threshold of 0.5. Thus the condition is expressed as in Equation 7:

$$\text{if } P(Y=1|X) \geq 0.5, \text{ predict } Y=1; \text{ otherwise, predict } Y=0. \tag{7}$$

3.5 SVM

Model Overview

The Support Vector Machine (SVM) is a supervised machine learning model primarily used for classification tasks. It is designed to find the optimal hyperplane that maximizes the margin between two classes. In this context, the margin is defined as the distance between the hyperplane and the nearest data points from each class, which are known as *support vectors*.

Linear SVM Formulation

For a binary classification problem, let $X = \{x_1, x_2, \ldots, x_n\}$ represent the set of input feature vectors, and $y = \{y_1, y_2, \ldots, y_n\}$ denote the corresponding labels, where $y_i \in \{-1, +1\}$ for each $i = 1, \ldots, n$. The goal is to find a hyperplane that can be defined by Equation 8 as shown below :

$$w^T x + b = 0 \tag{8}$$

where:

- w is the weight vector perpendicular to the hyperplane, and
- b is the bias term.

For a correctly classified point, the following constraints must hold the following Equation 9:

$$y_i(w^T x_i + b) \geq 1, \quad \forall i = 1, \ldots, n \tag{9}$$

This constraint ensures that data points from each class are separated by a margin of at least 1.

Optimization Problem

To maximize the margin, the objective is to minimize $\frac{1}{2}\|w\|^2$, subject to the constraints in Equation (2). The optimization problem is then formulated as Equation 10:

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 \tag{10}$$

$$\text{subject to} \quad y_i(w^T x_i + b) \geq 1, \quad \forall i = 1, \ldots, n \tag{11}$$

This is a convex optimization problem that can be solved using the method of Lagrange multipliers.

Lagrangian Dual Formulation

To solve the constrained optimization, the Lagrangian is constructed governed by Equation 12 :

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(w^T x_i + b) - 1 \right) \tag{12}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers.

The dual formulation, obtained by differentiating L with respect to w and b and setting the derivatives to zero, is given by:

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{13}$$

$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{14}$$

$$\alpha_i \geq 0, \quad \forall i = 1, \ldots, n \tag{15}$$

The weight vector w can then be expressed as a linear combination of the support vectors:

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{16}$$

Kernel Trick for Non-Linear SVM

For non-linearly separable data, the *kernel trick* is employed to project the data into a higher-dimensional space where a linear separation is possible. A kernel function $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is used, where $\phi(\cdot)$ is a mapping to a higher-dimensional space. Common kernel functions include:

- **Linear kernel:** $K(x_i, x_j) = x_i^T x_j$
- **Polynomial kernel:** $K(x_i, x_j) = (x_i^T x_j + 1)^d$
- **Gaussian (RBF) kernel:** $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$

The dual formulation is modified to use the kernel function as shown in Equation 17 :

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{17}$$

$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{18}$$

$$\alpha_i \geq 0, \quad \forall i = 1, \ldots, n \tag{19}$$

Soft Margin for Non-Separable Cases

In cases where data are not linearly separable even after kernel transformation, a *soft margin* SVM is applied by introducing slack variables $\xi_i \geq 0$ that allow some misclassifications. The optimization problem thus becomes as described inEquation 20 :

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \tag{20}$$

$$\text{subject to} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n \tag{21}$$

$$\xi_i \geq 0, \quad \forall i = 1, \ldots, n \tag{22}$$

where C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification errors.

Decision Function

Once w and b have been determined, the decision function for a new input x is given by Equation 23 as shown below

$$f(x) = w^T x + b \tag{23}$$

For kernelized SVMs, this becomes:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \tag{24}$$

The classification decision is made as following Equation 25 :

$$\text{if } f(x) \geq 0, \text{ predict } Y = +1; \text{ otherwise, predict } Y = -1. \tag{25}$$

3.6 Decision Tree

*Introduction A Decision Tree is a supervised learning algorithm used for both classification and regression tasks. In this model, data is continuously split according to a certain parameter until a specific criterion is met. The structure resembles a tree, where each internal node represents a feature-based decision, each branch represents an outcome of the decision, and each leaf node represents the final prediction or classification.

Model Structure and Split Criterion

A Decision Tree recursively partitions the data into subsets to maximize the homogeneity of target labels in each subset. The choice of split at each internal node is made based on a criterion that measures the quality of the split. Among the most commonly used criteria are Gini Impurity, Entropy, and Variance Reduction.

Gini Impurity (Classification)

For classification tasks, Gini Impurity is often used to evaluate the quality of a split. Gini Impurity measures the probability of misclassifying a randomly chosen element from the subset if it were labeled according to the distribution of labels in that subset.

Given a node t containing samples belonging to K classes, the Gini Impurity, $G(t)$, is defined as in Equation 26 :

$$G(t) = 1 - \sum_{k=1}^{K} p_k^2 \tag{26}$$

where p_k represents the proportion of samples in node t that belong to class k.

When a split is made, the Gini Impurity is calculated for each resulting subset, and the weighted average impurity is computed. The reduction in impurity, known as the Gini Gain, is then used to determine the optimal split.

Entropy and Information Gain (Classification)

Another criterion commonly used for classification is Entropy, which measures the impurity or disorder within a set of data. Entropy, $H(t)$, for a node t is defined as following Equation 27 :

$$H(t) = -\sum_{k=1}^{K} p_k \log_2 p_k \tag{27}$$

where p_k is the proportion of samples in node t belonging to class k.

To determine the effectiveness of a split, Information Gain is computed. Information Gain is defined as the reduction in entropy after the dataset is split according to a particular attribute. For a node t that is split into two child nodes t_L and t_R, the Information Gain IG is calculated as in Equation 28 :

$$IG = H(t) - \left(\frac{N_L}{N} H(t_L) + \frac{N_R}{N} H(t_R) \right) \tag{28}$$

where N is the number of samples in node t, and N_L and N_R are the number of samples in the left and right child nodes, respectively.

Variance Reduction (Regression)

For regression tasks, the split criterion aims to minimize the variance in each subset, as opposed to maximizing homogeneity of classes. The variance of a node t containing N samples with target values y_i is calculated as per following Equation 29 :

$$\mathrm{Var}(t) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{29}$$

where $\bar{y}$ represents the mean of target values in node t.

The variance reduction from a split into two nodes t_L and t_R is then calculated as shown in Equation 30 :

$$\Delta\mathrm{Var} = \mathrm{Var}(t) - \left(\frac{N_L}{N} \mathrm{Var}(t_L) + \frac{N_R}{N} \mathrm{Var}(t_R) \right) \tag{30}$$

where N, N_L, and N_R are the numbers of samples in nodes t, t_L, and t_R, respectively.

Tree Growth and Stopping Criteria

The Decision Tree is grown recursively by selecting the best split for each node based on the chosen criterion (e.g., Gini Impurity, Entropy, or Variance). The growth of the tree continues until one of the stopping criteria is met. Common stopping criteria include:

- A maximum depth limit is reached.
- A minimum number of samples per leaf node is reached.
- The decrease in impurity or variance falls below a specified threshold.

Pruning

Pruning is a technique used to reduce the size of the tree and improve generalization by preventing overfitting. Post-pruning is a commonly used approach, in which the fully grown tree is pruned by removing branches that contribute little to the predictive power of the model. Cost-complexity pruning (or *weakest link pruning*) involves defining a cost function that balances tree complexity and misclassification error as shown in Equation 31:

$$C_\alpha(T) = R(T) + \alpha \times |T| \tag{31}$$

where $R(T)$ is the misclassification rate of the tree T, $|T|$ is the number of terminal nodes in the tree, and α is a tuning parameter that controls the trade-off between model complexity and accuracy.

3.7 Random Forest

Overview of the Random Forest Model

Random Forest is an ensemble learning method primarily used for classification and regression tasks. This model consists of multiple decision trees, which together form a "forest." The predictions of the individual trees are aggregated to form the final prediction. Random Forest addresses the limitations of single decision trees, such as high variance, by averaging multiple decision trees created from bootstrapped samples of the dataset.

Construction of Decision Trees

Let $\{(X^{(i)}, Y^{(i)})\}_{i=1}^{N}$ denote a dataset with N samples, where each $X^{(i)} = (X_1^{(i)}, X_2^{(i)}, \ldots, X_n^{(i)})$ represents a feature vector and $Y^{(i)}$ denotes the corresponding label.

Each tree in the forest is trained on a bootstrapped sample of the original dataset. Given the bootstrapped sample D_t, the tree is grown by recursively partitioning the data at each node. At each node, a random subset of m features from the total n features is selected, and the optimal split among these m features is determined according to a chosen criterion, such as Gini impurity or entropy for classification.

Gini Impurity and Entropy for Splitting

To evaluate the quality of a split at each node, Gini impurity and entropy are commonly used metrics. For a node containing samples from K classes, let p_k denote the probability of a sample belonging to class k at this node. These metrics are defined as follows:
Gini Impurity The Gini impurity I_G at a node is calculated as shown in following Equation 32 :

$$I_G = 1 - \sum_{k=1}^{K} p_k^2 \tag{32}$$

Entropy The entropy H at a node is given by Equation 33:

$$H = -\sum_{k=1}^{K} p_k \log_2(p_k) \tag{33}$$

For each possible split, these metrics are calculated to identify the one that results in the greatest reduction in impurity or entropy.

Aggregation of Predictions

The final prediction in a Random Forest model is obtained by aggregating the predictions of all individual trees in the forest.
Classification In classification tasks, the final predicted class as shoen in Equation 34 $\hat{Y}$ is determined by a majority vote among the predictions $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_T$ from the T trees:

$$\hat{Y} = \text{mode}(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_T) \tag{34}$$

Regression: For regression tasks, the final prediction as shoen in Equation 35 $\hat{Y}$ is obtained by averaging the individual predictions from each tree:

$$\hat{Y} = \frac{1}{T} \sum_{t=1}^{T} \hat{Y}_t \tag{35}$$

Out-of-Bag Error

In Random Forest, out-of-bag (OOB) error estimation is used as an internal mechanism to assess model accuracy without the need for a separate validation set. Given a sample that is not included in the bootstrapped sample for a particular tree t, this sample can serve as a test instance for that tree. The OOB error is calculated by averaging the prediction errors over all trees for which the sample was not included in the training set.

Feature Importance

Feature importance scores provide insight into the contribution of each feature in making predictions. In Random Forest, the importance of feature X_j can be computed by measuring the average reduction in Gini impurity (or entropy) across all nodes that split on X_j across all trees. The feature importance score for X_j, denoted $I(X_j)$, is given by Equation 36:

$$I(X_j) = \frac{1}{T} \sum_{t=1}^{T} \sum_{n \in N_t} \Delta I_{n,j} \tag{36}$$

where:

- T is the total number of trees,
- N_t is the set of nodes in tree t,
- $\Delta I_{n,j}$ represents the reduction in impurity at node n by splitting on feature X_j.

3.8 Gradient Boosting

Model Overview

Gradient Boosting is an ensemble learning method for regression and classification tasks. This model builds a sequence of weak learners, typically decision trees, in a stage-wise manner, such that each subsequent model attempts to correct the errors made by its predecessor. The primary goal is to minimize a predefined loss function by iteratively adding models that reduce residual errors.

Functional Gradient Descent

The Gradient Boosting method can be understood as an application of functional gradient descent, where a sequence of models is constructed to approximate a function $F(x)$ that minimizes a given loss $L(y, F(x))$. The objective is expressed as Equation 37 that follows as:

$$F^*(x) = \arg\min_F \mathbb{E}_{x,y}\big[L(y, F(x))\big] \tag{37}$$

where y represents the actual target variable, x denotes the input features, and $F(x)$ is the model's prediction.

An initial model $F_0(x)$ is chosen to approximate $F(x)$. Then, subsequent models are added to minimize the residual errors iteratively.

Additive Model Formulation

Gradient Boosting employs an additive model, defined as Equation 38 :

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{38}$$

where:

- $F_{m-1}(x)$ represents the prediction from the previous stage,
- $h_m(x)$ is the new model (or weak learner) added at stage m,
- γ_m is the step size or learning rate.

The new model $h_m(x)$ is selected to minimize the residual error between the true value and the prediction from the previous stage.

Loss Function and Residuals

The loss function $L(y, F(x))$ guides the selection of the next model. For each observation i at stage m, the residual r_{im} as defined in Equation 39, is defined as the negative gradient of the loss function with respect to the prediction:

$$r_{im} = -\left.\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right|_{F(x)=F_{m-1}(x)} \tag{39}$$

This residual r_{im} can be interpreted as the direction and magnitude by which the current prediction $F_{m-1}(x)$ deviates from minimizing the loss function for each observation.

Fitting the Weak Learner

At each stage m, a new model $h_m(x)$ is trained to predict the residuals r_{im} computed for each observation. This model is fitted to approximate the residuals in order to improve the overall prediction.

Update Rule

The weak learner $h_m(x)$ is scaled by a factor γ_m, known as the step size or learning rate, to control the contribution of each stage to the final model. The update rule for the model at stage m can then be expressed as Equation 40 :

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{40}$$

Typically, γ_m is a fixed learning rate that is determined through tuning and remains constant across iterations.

Table 2: Summary of the dataset features

Feature	Datatype	Mean	Standard Deviation
Age	Int64	53.60	8.732
Urea	Float64	5.123	2.940
Creatinine Ratio	Int64	68.937	60.131
HbA1C	Float64	8.284	2.534
Cholesterol	Float64	4.865	1.303
Triglycerides	Float64	2.352	1.403
HDL	Float64	1.205	0.661
LDL	Float64	2.609	1.117
VLDL	Float64	1.824	3.621
BMI	Float64	29.561	4.952

Convergence and Regularization

The convergence of the Gradient Boosting model can be controlled by setting the maximum number of stages M, the learning rate γ_m, and the complexity of each weak learner $h_m(x)$. Regularization techniques, such as limiting the depth of decision trees, early stopping, and subsampling, are often employed to prevent overfitting and improve generalization.

4 Experiment

4.1 Objective of the experiment

Based on the problem statement described in previous section, the objective of the experiment could be identified as

- To implement the selected learning models on to a standard and benchmarked dataset containing identified features.
- To record performance metrics of all the selected models to have a concrete conclusion on which model should be chosen for computational diabetic decision making.
- To explain the experimental observations through theoretical knowledge.

4.2 Dataset Description

The dataset used in this study has been collected from *https://data.mendeley.com/datasets/wj9rwkp9c2/1*. This dataset complies crucial medical and lab data of 1,000 patients of Medical City Hospital and Al-Kindy Teaching Hospital in Iraq. consists of medical records of patients, which include the Medical City Hospital and Al-Kindy Teaching Hospital in Iraq.

The dataset contains all the attributes mentioned in subsection 3.1. The mapping of the attributes to the notations mentioned in subsection 3.1 are as follows.

Attribute *No. of Patient*, *Sugar Level (Blood Glucose)*, *Age Gender*, *Creatinine Ratio*, *Body Mass Index*, *Urea*, *Cholesterol* and *HbA1C* directly map to *pat_id*, *sl*, *ag*, *gnd*, *cr*, *bmi*, *ur*, *ch* and *hba1c* respectively. The symbol *flp* designates a vector of LDL (Low-Density Lipoprotein), VLDL (Very Low-Density Lipoprotein), Triglycerides (TG) and HDL (High-Density Lipoprotein) attributes of the dataset.

The distribution of key features in the dataset is summarized in Table 2.

4.3 Data Preparation

The dataset underwent several preprocessing steps to ensure its quality for training the machine learning models:

Label Encoding

The categorical columns in the dataset, specifically *Gender* and *Class*, were encoded into numerical values using label encoding. Gender was coded as 0 for Male and 1 for Female, while the *Class* column was encoded as 0 for Non-Diabetic, 1 for Predict-Diabetic, and 2 for Diabetic.

Outlier Detection and Removal

Outliers were detected using the Interquartile Range (IQR) method. This method calculates the spread of the middle 50% of the data, and outliers are identified as values that fall outside 1.5 times the IQR. These outliers were removed to improve model performance.

Feature Standardization

The numerical features were standardized to ensure all variables were on the same scale, with a mean of 0 and a standard deviation of 1. Standardization is crucial for models like Support Vector Machines and Gradient Boosting, which are sensitive to the scale of input features.

Train-Test Split

The dataset was split into training (70%) and testing (30%) sets to evaluate the model's generalization performance. This split was chosen to ensure a sufficient amount of data for both model training and evaluation.

4.4 Model Implementation

Hyperparameter Tuning

Each model underwent hyperparameter tuning to optimize its performance. The hyperparameter tuning was done using `GridSearchCV`, which systematically searches for the best parameter values based on model performance on a validation set.

Model Evaluation

The performance of the models was evaluated using several metrics, including classification accuracy, precision, recall, F1-score, and the confusion matrix. These metrics provide a comprehensive view of the models' ability to correctly identify diabetic, non-diabetic, and predict-diabetic cases.

Classification Accuracy

Accuracy measures the ratio of correct predictions to the total number of predictions made.

Confusion Matrix

The confusion matrix provides detailed insights into the model's performance, showing the number of true positives, false positives, true negatives, and false negatives for each class.

Precision, Recall, and F1-Score

- Precision: The proportion of true positive predictions out of all predicted positive cases.
- Recall: The proportion of true positives out of all actual positives.
- F1-Score: The harmonic mean of precision and recall, balancing the two metrics.

Table 3: Model performance comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	89.4%	0.75	0.69	0.72
SVM	94.1%	0.77	0.77	0.77
Decision Tree	95.8%	0.91	0.77	0.83
Random Forest	98.22%	1.00	0.92	0.96
Gradient Boosting	97.6%	1.00	0.85	0.92

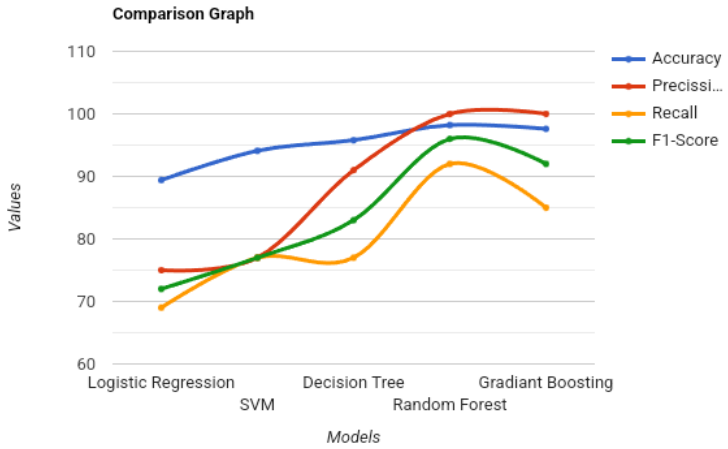

Figure 1: Comparisons of performance among ML algorithms.

ROC Curve and AUC

The ROC curve plots the true positive rate against the false positive rate, while the AUC score measures the model's ability to discriminate between classes.

4.5 Experimental Results

Among the implemented models, Random Forest achieved the highest accuracy at 98.22%. The performance of other models like SVM and Gradient Boosting was also strong, with accuracies exceeding 94%. A detailed comparison of the models based on accuracy, precision, recall, and F1-score is presented in Table 3.

Fig 1 shows the graphical representation of comparative study of model performance on chosen evaluation parameters. The graphical representation clearly shows the best performers on four different parameters.

4.6 Result Analysis

The experimental results, summarized in Table 3, demonstrate that the Random Forest model achieved the highest performance across all evaluated metrics, with an accuracy of 98.22%. Its precision, recall, and F1-score metrics also surpassed those of other models, highlighting its robustness and suitability for multiclass diabetes classification.

Among the other models, Gradient Boosting achieved the second-highest accuracy at 97.6%, with an impressive F1-score of 0.92, indicating its ability to balance precision and recall effectively. This performance validates the strength of ensemble methods in handling complex datasets.

The Decision Tree model performed moderately well, achieving an accuracy of 95.8%. While its precision was high at 0.91, the recall value was comparatively lower, indicating that it was slightly less effective in identifying all relevant cases compared to the Random Forest and Gradient Boosting models.

The SVM model, with an accuracy of 94.1%, displayed consistent performance across precision, recall, and F1-score, all at 0.77. This highlights its utility as a reliable classifier for balanced datasets but indicates a limitation in handling more complex patterns compared to ensemble methods.

Logistic Regression, as expected, showed the lowest accuracy at 89.4%, primarily due to its linear nature and inability to capture non-linear relationships inherent in the data. Its F1-score of 0.72 reflects moderate performance, suitable as a baseline but insufficient for precise multiclass classification.

Overall, the superior performance of Random Forest underscores the advantage of ensemble techniques in achieving high predictive accuracy and generalization. The model's ability to effectively handle feature importance and mitigate overfitting makes it the optimal choice for practical implementations in diabetes prediction tasks.

The observed results align with the theoretical understanding of diabetes pathophysiology and the characteristics of the chosen machine learning models. Random Forest, an ensemble method, excels in capturing complex interactions between features such as blood glucose levels, BMI, cholesterol, and HbA1c, which are crucial physiological indicators for diabetes classification. These indicators exhibit non-linear relationships and interdependencies, such as the impact of obesity on insulin resistance or the association between cholesterol and cardiovascular risks in diabetes. Random Forest's capability to aggregate decisions from multiple trees enhances its robustness and generalizability, particularly for datasets with diverse feature distributions. Similarly, Gradient Boosting's iterative refinement of residual errors complements the subtle nuances of diabetes progression, such as the transition from pre-diabetic to diabetic stages. On the other hand, models like Logistic Regression and SVM, which rely on linear separability or fixed decision boundaries, struggle to capture the multifaceted an

5 Conclusion

This study explored the application of machine learning algorithms for the multiclass classification of diabetes, distinguishing between diabetic, pre-diabetic, and non-diabetic individuals. By leveraging an extensive set of features, including physiological and biochemical indicators such as blood glucose levels, BMI, and lipid profiles, we developed models that achieved high accuracy. Among the models tested, Random Forest emerged as the best-performing algorithm, achieving an accuracy of 98.22%, followed closely by Gradient Boosting. These results emphasize the efficacy of ensemble learning techniques in handling complex datasets and providing reliable predictions in healthcare scenarios.

The findings validate the importance of feature diversity and advanced model architectures in capturing the multifaceted nature of diabetes progression. While simpler models like Logistic Regression provide baseline results, their inability to model non-linear relationships limits their applicability in this context. Conversely, ensemble methods demonstrate robustness and generalization, making them suitable for real-world deployments in clinical decision-making.

Future research can build upon this work by integrating additional genetic and behavioral data to further enhance prediction accuracy. Additionally, designing a novel activation function tailored to the underlying science of diabetes could be a promising avenue. Such an approach could improve the learning efficiency of neural networks, enabling them to better capture the intricate relationships between features and the progression of diabetes. This direction not only aligns with advancements in machine learning but also holds the potential to bridge the gap between computational techniques and medical science.

Authorship contribution statement

Soumen Ghosh: Conceptualization, Data curation, Formal analysis, Methodology; **Souvik Halder:** Experimentation, Validation, Visualization; **Prasun Maity:** Experimentation, Validation, Visualization; **Shiladitya Munshi:** Conceptualization, Writing – original draft, Writing – review and editing.

Conflict of Interest

The authors declare that there is no conflict of interest in this work.

Data availability

Data may be available on request.

Acknowledgments

This work is fully sponsored by KinetiCraft Career Solutions Private Limited (https://www.kineticraft.in/) under their R&D funding scheme.

References

Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A. T., Ghazal, T. M., & Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. *IEEE Access*, *10*, 8529–8538. https://doi.org/10.1109/ACCESS.2022.3142097.

Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., Hussain, A., Malik, M. A., Raza, M. M., Ibrar, S., et al. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, 16, 100204. https://doi.org/10.1016/j.imu.2019.100204.

Anjana, R. M., Unnikrishnan, R., Deepa, M., Pradeepa, R., ..., & Mohan, V. (2023). Metabolic non-communicable disease health report of india: the icmr-indiab national cross-sectional study (icmr-indiab-17). *The Lancet Diabetes & Endocrinology*, *11*(7), 474–489. https://doi.org/10.1016/S2213-8587(23)00119-5.

Ayon, S. I. & Islam, M. M. (2019). Diabetes prediction: a deep learning approach. *International Journal of Information Engineering and Electronic Business*, 13(2), 21. https://doi.org/10.5815/ijieeb.2019.02.03.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Bukhari, M. M., Alkhamees, B. F., Hussain, S., Gumaei, A., Assiri, A., & Ullah, S. S. (2021). An improved artificial neural network model for effective diabetes prediction. *Complexity*, *2021*(1), 5525271. https://doi.org/10.1155/2021/5525271.

Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine learning based diabetes classification and prediction for healthcare applications. *Journal of Healthcare Engineering*, *2021*, 9930985. https://doi.org/10.1155/2021/9930985.

Carstensen, D. H. et al. (2020). Age as a determinant in the risk of developing type 2 diabetes: A population-based cohort study. *Diabetes Care*, *35*(5), 999–1003.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785.

Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964.

Davis, M. & Lee, K. (2019). Logistic regression as a baseline for predictive modeling. *Statistical Methods in Health Research*, *10*(1), 10–18.

Efron, B. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Fox, J. A. & Flegal, M. T. (2023). Bmi as a risk factor for diabetes incidence. *International Journal of Obesity*, *37*, 1019–1027.

Fruchart, J. H. et al. (2021). Significance of fasting lipid profile in diabetes risk assessment. *Diabetes Metabolism*, *46*(4), 289–297.

Gerstein, M. N. & Ong, A. Y. (2024). Cholesterol and cardiovascular risk in diabetes. *Journal of the American College of Cardiology*, *71*(19), 2330–2340.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://doi.org/10.1007/s10710-017-9314-z.

Guyon, I. & Gunn, K. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182. https://doi.org/10.1162/153244303322753616.

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516–76531. https://doi.org/10.1109/ACCESS.2020.2989857.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7.

He, S. H. & Zhang, Y. (2023). A comprehensive survey on gradient boosting machines. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2100–2113.

Huang, Y. & Li, T. (2021). Robustness of support vector machines in high-dimensional spaces. *Pattern Recognition Letters*, *35*, 200–206. https://doi.org/10.1016/j.patcog.2024.110544.

Jayanthi, N., Babu, B. V., & Rao, N. S. (2017). Survey on clinical prediction models for diabetes prediction. *Journal of Big Data*, *4*(1). https://doi.org/10.1186/s40537-017-0082-7.

Khanam, J. J. & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, *7*(4), 432–439. https://doi.org/10.1016/j.icte.2021.02.004.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539.

Liang, X., Alshemmary, E. N., Ma, M., Liao, S., Zhou, W., & Lu, Z. (2021). Automatic diabetic foot prediction through fundus images by radiomics features. *IEEE Access*, *9*, 92776–92787. https://doi.org/10.1109/ACCESS.2021.3093358.

Liaw, A. & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22.

Mujumdar, A. & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, 292–299. https://doi.org/10.1016/j.procs.2020.01.047.

Nai-Arun, N. & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science*, 69, 132–142. https://doi.org/10.1016/j.procs.2015.10.014.

Qiao, L., Zhu, Y., & Zhou, H. (2020). Diabetic retinopathy detection using prognosis of microaneurysm and early diagnosis system for non-proliferative diabetic retinopathy based on deep learning algorithms. *IEEE Access*, *8*, 104292–104302. https://doi.org/10.1109/ACCESS.2020.2993937.

Quinlan, R. (2014). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Rao, B. G. N. & Narasimhan, L. R. (2003). Naive bayes vs. decision trees. *Journal of Statistical Computation and Simulation*, *73*(4), 397–412.

Rennie, N. (1997). On the bias-variance tradeoff for linear and naive bayes classifiers. In *Proceedings of the 14th International Conference on Machine Learning* (pp. 299–306).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778.

Shi, C. & Zhang, Z. (2021). Handling missing data in gradient boosting. *Data Mining and Knowledge Discovery*, *31*(4), 1023–1039. https://doi.org/10.1088/1742-6596/1684/1/012062.

Sinha, R. W. & Lipton, R. M. (2021). Glucose levels and diabetes risk. *Diabetes Care*, *27*, 573–582.

Smith, J. & Johnson, A. (2021). Application of logistic regression in predicting diabetes. *Journal of Medical Informatics*, *45*(2), 123–130.

Vapnik, P. (2013). *The Nature of Statistical Learning Theory*. Springer. https://doi.org/10.1016/j.patcog.2024.110544.

Wilson, J. D. & Porter, M. C. (2022). Glycated hemoglobin as a predictive marker for diabetes. *The Lancet Diabetes Endocrinology*, *9*(3), 204–212. https://doi.org/10.1093/ajcn/nqac154.

Xue, J., Min, F., & Ma, F. (2020). Research on diabetes prediction method based on machine learning. *Journal of Physics: Conference Series*, *1684*, 012062. https://doi.org/10.1088/1742-6596/1684/1/012062.