Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1618267

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 8 - Issue 3: 875-884/ May 2025

PERFORMANCE EVALUATION OF SUPERVISED MACHINE LEARNING CLASSIFIERS FOR TYPE 2 DIABETES MELLITUS PREDICTION

Onur KURT[1]*

[1]Istanbul Technical University, Faculty of Electrical and Electronics, Department of Electronics and Communication Engineering, 34469, Istanbul, Türkiye

Abstract: Diabetes mellitus is a significant global health concern that profoundly affects individuals' lives and imposes a considerable burden on healthcare systems. Enhanced predictive capabilities can lead to timely interventions, ultimately improving patient outcomes and alleviating the strain on healthcare resources. Thus, accurate and timely prediction of diabetes mellitus is crucial for reducing mortality rates and minimizing complications within healthcare frameworks. This study addresses the correlation between type 2 diabetes mellitus (T2DM) and key attributes that differentiate diabetic from non-diabetic cases, utilizing various machine learning-based classification methods. For this reason, this work employed a large, open-source dataset obtained from Kaggle. To my knowledge, this is the first study utilizing such a dataset that specifically focuses on predicting T2DM in patients aged 35 years or older, according to the American Diabetes Association (ADA). To identify key features associated with T2DM for use as input to each supervised classifier, the Minimum Redundancy Maximum Relevance (mRMR) feature selection algorithm was applied to the dataset. In this analysis, the performance of each supervised classifier with feature selection was evaluated and compared using various metrics, including accuracy, sensitivity, specificity, precision (positive predictive value, PPV), negative predictive value (NPV), F1 score, and the area under the receiver operating characteristic curve (AUROC). The results of the analysis reveal that an ensemble method employing boosted trees (EBT) classifier surpasses the other models, recording the highest macro-average values for accuracy (95.9%), PPV (97.7%), NPV (97.7%), and F1 score (89.7%), along with the superior area under the curve (AUC) of 95.57% for both diabetes and non-diabetes cases. The study suggests that machine learning classifiers can serve as a reliable tool for the precise prediction of T2DM, thereby enhancing clinical decision-making processes for healthcare practitioners.

Keywords: Classification, Feature selection, Supervised learning, Type 2 diabetes mellitus

***Corresponding author:** Istanbul Technical University, Faculty of Electrical and Electronics, Department of Electronics and Communication Engineering, 34469, Istanbul, Türkiye
E mail: onurkurt@itu.edu.tr (O. KURT)
Onur KURT https://orcid.org/0000-0002-4486-2257

Received: January 12, 2025
Accepted: April 22, 2025
Published: May 15, 2025

Cite as: Kurt O. 2025. Performance evaluation of supervised machine learning classifiers for type 2 diabetes mellitus prediction. BSJ Eng Sci, 8(3): 875-884.

1. Introduction

Diabetes mellitus is a major health challenge that significantly affects individuals on a global scale (Lin et al., 2020). Diabetes mellitus is a chronic condition marked by elevated blood glucose levels, referred to as hyperglycemia. It occurs when the body either produces insufficient insulin, a hormone secreted by the pancreas, or is unable to utilize the insulin effectively (WHO, 2024). Over time, diabetes significantly increases the risk of damage to the eyes, kidneys, nerves, and heart, potentially leading to severe complications such as blindness, heart disease, stroke, and kidney failure (Zhou et al., 2023). The International Diabetes Federation (IDF) reports that approximately 537 million individuals globally are affected by diabetes mellitus in 2021, with projections indicating an increase to 643 million by 2030 and 783 million by 2045 (IDF, 2021). Diabetes mellitus can be broadly classified into three types: Type 1 diabetes mellitus (T1DM), Type 2 diabetes mellitus (T2DM), and gestational diabetes mellitus (GDM) (Rastogi and Bansal, 2023). T1DM, a complex autoimmune disorder characterized by the destruction of insulin-producing pancreatic β cells, is most commonly diagnosed in children (Cano-Cano et al., 2022; Costa-Cordella et al., 2021). In contrast, T2DM, also known as insulin-independent diabetes, arises due to insulin resistance or insufficient insulin production and is predominantly diagnosed in middle-aged and older adults (Ma et al., 2022; Varma et al., 2014; Carrillo-Larco et al., 2024). GDM, the third major form of diabetes, typically develops during pregnancy (Fazakis et al., 2021). Although each type of diabetes impacts individual health and places a significant burden on healthcare systems worldwide, over 90% of all diabetes cases are attributed to T2DM, making it the most prevalent form of the disease globally (Borse et al., 2021; Laakso and Kuusisto, 2014). According to the World Health Organization (WHO), the global prevalence of diabetes is projected to rise substantially, leading to an increase in mortality rates worldwide (WHO, 2024). Thus, early detection and effective management of diabetes are essential not only for improving individual health outcomes but also for mitigating complications and reducing the burden on healthcare systems. Recently, machine learning (ML) algorithms have gained significant popularity, particularly in engineering and science, and are now employed in a wide range of applications, including medical diagnostics, computer vision, predictive analytics in environmental science, and image recognition, among others. ML is a subset of artificial intelligence (AI) that facilitates the training of computers to make precise predictions based on provided input data (Janiesch et al., 2021; Jordan and Mitchell, 2015; Kurt, 2024). In recent years, significant efforts have focused on ML techniques to predict diabetes mellitus, offering advantages over conventional diagnostic methods. Rastogi and Bansal (2023) proposed data mining techniques to predict diabetes based on dataset collected from Kaggle. The authors employed four data mining techniques, namely Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and Naive Bayes (NB), and compared their performance using accuracy and sensitivity. Their findings showed that the highest accuracy of 82.46% was achieved with LR compared to the other models. Febrian et al. (2023) conducted a study to predict diabetes using two ML models, namely k-nearest neighbors (KNN) and NB, based on the Pima Indians Diabetes dataset obtained from Kaggle. Their study indicated that NB outperforms KNN, with an average accuracy of 76.07%, an average precision of 73.37%, and an average recall of 71.37%. Tasin et al. (2022) conducted a comparative analysis of various ML models for predicting diabetes, using the open-source Pima Indians dataset and a private dataset of female Bangladeshi patients. In their study, they also utilized SMOTE and ADASYN preprocessing techniques to handle the issue of imbalanced class problems. Their results demonstrated that the Extreme Gradient Boosting (XGBoost) classifier achieved the best performance with 81% accuracy and an F1 score and the area under the curve (AUC) of 0.81 and 0.84, respectively, with the ADASYN approach. Additionally, Tigga and Garg (2020) implemented six ML models and compared their results to obtain a suitable model for predicting T2DM using a dataset collected through both online and offline questionnaires. The same algorithms were also applied to the Pima Indian Diabetes dataset. Their experimental results indicated that the accuracy of the RF model is 94.10%, the highest among the other models. Similarly, Talukder et al. (2024) utilized several datasets obtained from publicly available sources. They utilized a range of ML models to predict diabetes mellitus using four different datasets. The performance analysis revealed that, among all ML algorithms, RF outperforms existing methods with accuracy rates of 86% and 98.48% for Dataset 1 and Dataset 2, respectively, while XGBoost and Decision Tree (DT) algorithms achieve accuracy rates of 99.27% and 100% for Dataset 3 and Dataset 4, respectively. In addition, Modak and Jha (2024) estimated diabetes using the Diabetic2 dataset collected from Kaggle by employing various ML models. They demonstrated that CatBoost stands out as the most effective model, achieving an impressive accuracy rate of 95.4%, surpassing XGBoost's accuracy of 94.3%. Futhermore, CatBoost's superior AUC-ROC score of 0.99 further underscores its potential advantage over XGBoost, which attained an AUC-ROC score of 0.98. Bhat et al. (2023) developed three ML-based classification algorithms: LR, Gradient Boosting (GB), and DT, combined with a feature selection method, to forecast different types of diabetes using the Pima Indian Diabetes Dataset. They obtained the highest accuracy rate (91 %), precision (96 %), recall (92 %), and F1 score (94 %) using DT. Similarly, Zhou et al. (2023) conducted a study to predict diabetes utilizing Boruta feature selection and ensemble learning with the Pima Indian Diabetes Dataset. The obtained results showed that the model achieves an accuracy rate of 98% and demonstrates strong performance. While these studies offer significant contributions to the prediction of diabetes mellitus, they mainly focused on the Pima Indians dataset, either in isolation or in combination with other datasets. Therefore, the desire to enhance the accuracy of ML-based predictions of diabetes necessitates a detailed study based on a different dataset. This paper reports a comprehensive study on classification-based prediction of diabetes mellitus using various ML models applied to a multi-feature diabetes prediction dataset obtained from Kaggle. To attain this target, the study utilizes a range of ML models, including linear discriminant analysis (LDA), LR, Gaussian NB, fine Gaussian SVM, Fine KNN, and an ensemble method employing boosted trees (EBT). The choice of these models for diabetes prediction was guided by their strengths and appropriateness for the task. Furthermore, Minimum Redundancy Maximum Relevance (mRMR) feature selection method was utilized to identify key determinants associated with diabetes mellitus by analyzing medical history and demographic information. The results of the study reveal that EBT classifier stands out as the most reliable model among the other models for predicting diabetes based on the performance evaluation criteria. The findings suggest that the proposed models enhance the understanding of diabetes classification within the scientific community and may assist medical professionals in diabetes treatment and decision-making processes. To the best of my knowledge, this is the first study to conduct a detailed analysis using this specific dataset to predict T2DM in patients aged 35 years or older, as defined by the American Diabetes Association (ADA, 2021). This dataset, which has not been previously explored in the context of T2DM prediction for this age group, offers unique demographic, clinical, and possibly lifestyle-related variables that distinguish it from commonly used datasets in diabetes research. Unlike prior studies that rely on more generalized or widely available datasets, this research leverages a novel data source that enables a more precise and tailored predictive modeling approach. By doing so, it contributes to the existing literature by filling a critical gap in age-specific diabetes risk assessment, potentially leading to improved early detection and personalized intervention strategies. The structure of this paper is as follows: The introduction outlines the objectives of the study and includes a review of relevant literature. Section 2 presents an overview of the methods employed for preparing the dataset for ML models. Section 3 discusses the results of the analysis, which are further elaborated in Section 4 through a comprehensive discussion. Finally, the conclusion and potential future directions are outlined in Section 5.

2. Materials and Methods

2.1. Data Description

The present study utilized an open-source dataset, referred to as the "diabetes prediction dataset", which was collected from Kaggle (Kaggle, 2024). The dataset comprises 8 features that are associated with diabetes, which serves as the target variable for analysis. These features (gender, age, hypertension, heart disease, smoking history, body mass index, HbA1c level, blood glucose level) include various clinical and demographic attributes that may influence the prediction of diabetes outcomes. The target variable in this study is the diabetes status of patients, classified as either positive (indicating the presence of diabetes) or negative (indicating the absence of diabetes). This study employed binary classification, designating the outcome variable as '0' for non-diabetic patients and '1' for diabetic patients, with diabetes status serving as the primary outcome variable. Descriptions of all variables included in the analysis is presented in Table 1. Since T2DM accounts for over 90% of all diabetes cases (Borse et al., 2021; Laakso and Kuusisto, 2014), the primary focus of the current study is based on data related to T2DM. According to the American Diabetes Association (ADA), the recommended age of annual diabetes screening tests for T2DM has decreased to as young as 35 years (ADA, 2021). For this reason, this study only considers patients who are 35 years of age or older. The statistical values of all numerical features, as well as the distributions of all categorical and nominal features, including the target variable, are presented in Table 2 and Table 3, respectively.

Table 1. Descriptions of all variables including numerical, categorical, and nominal

Variable Name	Type	Description
Gender	Categorical	Gender of the individual (0 = Female, 1 = Male)
Age	Continuous	Age of the individual in years
Hypertension	Categorical	Presence of hypertension (0 = No, 1 = Yes)
Heart Disease	Categorical	Presence of heart disease (0 = No, 1 = Yes)
Smoking History Never	Nominal	Smoking history - Never smoked (0 = No, 1 = Yes)
Smoking History Former	Nominal	Smoking history - Former smoker (0 = No, 1 = Yes)
Smoking History Not Current	Nominal	Smoking history - Not current smoker (0 = No, 1 = Yes)
Smoking History Current	Nominal	Smoking history - Current smoker (0 = No, 1 = Yes)
Smoking History Ever	Nominal	Smoking history - Ever smoked (0 = No, 1 = Yes)
Smoking History No Info	Nominal	Smoking history - No information (0 = No, 1 = Yes)
BMI	Continuous	Body Mass Index
HbA1c Level	Continuous	Level of HbA1c - hemoglobin A1c test (a marker for diabetes management)
Blood Glucose Level	Continuous	Blood glucose level measured in mg/dL
Diabetes	Categorical	Presence of diabetes (0 = No, 1 = Yes)

Table 2. Summary statistics of all numerical features

Variable Name	Minimum	Maximum	Mean	Median	Mode	Standard Deviation
Age	35	80	56.873	56	80	13.547
BMI	10.01	91.82	29.025	27.32	27.32	6.042
HbA1c Level	3.5	9	5.598	5.8	5.7	1.114
Blood Glucose Level	80	300	140.966	145	130	43.671

Table 3. Distribution of all categorical and nominal features including target variable

Variable name	Count 0 (Non diabetes)	Count 1 (Diabetes)
Gender	36180	25053
Hypertension	54007	7226
Heart Disease	57315	3918
Smoking History Never	38532	22701
Smoking History Former	53034	8199
Smoking History Not Current	56685	4548
Smoking History Current	54875	6358
Smoking History Ever	58215	3018
Smoking History No Info	44824	16409
Diabetes	53172	8061

2.2. Data Preprocessing

Data preprocessing is a crucial step for preparing the dataset in an appropriate format prior to its use as input for ML classifiers. Data preprocessing was conducted using MATLAB R2022b. The flowchart illustrating the proposed methodology for diabetes mellitus prediction is presented in Figure 1. A subset of the dataset was generated by extracting specific data related to T2DM from the original dataset. As most ML algorithms are designed to work with numerical data, categorical data, such as gender, were converted into binary numerical format using label encoding, while smoking history was converted into binary numerical format using one-hot encoding. Since hypertension, heart disease, and diabetes are already labeled as binary numbers in the original dataset, no conversion was performed for those variables. In order to ensure that all features are on the same scale, continuous numerical features (age, body mass index, HbA1c level, blood glucose level) were normalized to a range between 0 and 1 using min-max scaling, as expressed by the following formula (equation 1):

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

where x_{norm} is the normalized value, x represents the original value in the dataset, and x_{min} and x_{max} represent the minimum and maximum values in the dataset, respectively.

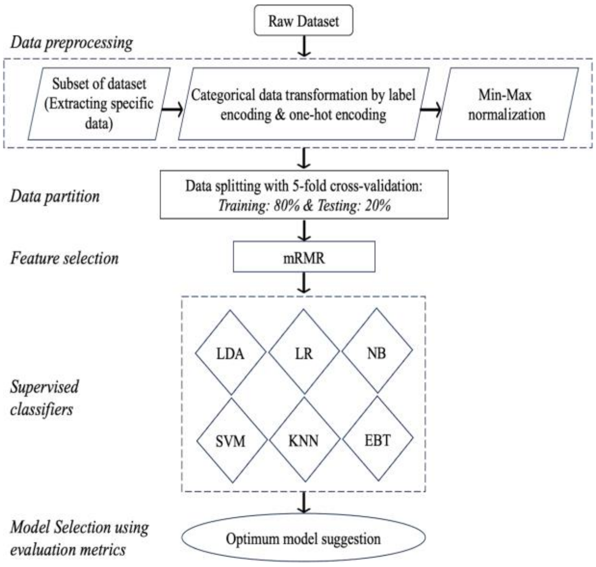
Raw Dataset
Data preprocessing
Subset of dataset (Extracting specific data)
Categorical data transformation by label encoding & one-hot encoding
Min-Max normalization
Data partition
Data splitting with 5-fold cross-validation: Training: 80% & Testing: 20%
Feature selection
mRMR
Supervised classifiers
LDA
LR
NB
SVM
KNN
EBT
Model Selection using evaluation metrics
Optimum model suggestion

Figure 1. Flow chat of the proposed methodology for type 2 diabetes mellitus prediction.

To train and evaluate the performance of all ML models for diabetes prediction, the dataset was split into training and testing sets with an 80:20 ratio. To prevent the risk of overfitting, all ML classifiers were trained using 5-fold cross-validation, and their performance was evaluated on a separate test dataset. Additionally, a feature selection method was employed to identify key attributes most relevant to diabetes mellitus within the dataset. The mRMR feature selection method, which effectively handles both numerical and categorical variables, was selected to extract a subset of features from the dataset. The selection of the mRMR feature selection method for the dataset with mixed data types is justified by several key factors. Primarily, mRMR's ability to handle both continuous and categorical variables through mutual information computation makes it particularly suitable for such diverse dataset. The method's dual optimization approach, which maximizes feature relevance while minimizing redundancy, ensures a more informative and diverse feature subset, potentially enhancing model performance and interpretability. Additionally, mRMR's computational efficiency and scalability to larger datasets make it a practical choice for high-dimensional data analysis. Its model-agnostic nature allows for flexibility in subsequent modeling approaches, enabling experimentation with various algorithms. Furthermore, mRMR has demonstrated effectiveness across multiple domains, including bioinformatics, finance, and healthcare, lending empirical support to its reliability in selecting features that significantly contribute to predictive accuracy. In comparison to alternative methods such as filter methods (e.g., chi-square, ANOVA), wrapper methods (e.g., forward selection), and embedded methods (e.g., LASSO), mRMR offers a balanced approach that avoids the pitfalls of independent feature evaluation, computational expense, overfitting, and model-specificity. Given these considerations, mRMR presents the optimal balance of theoretical rigor and practical utility for this study, aligning well with the characteristics of the dataset and the requirements of this analysis. The mRMR can be mathematically represented as (equation 2):

$$R(x_i, c) = I(x_i; c) \tag{2}$$

where R represents relevance, x_i is a feature, c is the target variable, and $I(x_i;c)$ denotes the mutual information between feature x_i and the target variable c (equation 3).

$$R(S) = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j) \tag{3}$$

where R represents redundancy among selected features, S is the set of selected features, and $I(x_i;x_j)$ is the mutual information between x_i and x_j.

By combining relevance and redundancy, the mRMR becomes as follows (equation4):

$$mRMR(x_i) = R(x_i, c) - R(S) \tag{4}$$

Figure 2 presents the normalized importance scores of individual features as determined by the mRMR feature selection method. In the feature selection process, a moderate threshold value of 0.12 was selected based on the distribution of normalized importance scores across all features, as shown in Figure 2. This threshold was determined through iterative testing across multiple supervised classifiers, where cross-validation experiments with varying threshold values indicated that 0.12 achieved an optimal balance between model complexity and predictive performance. This threshold retained the top five features while excluding those with minimal contribution to model accuracy. The five significant diabetes-related features identified through the mRMR method are blood glucose level, smoking history ever, hypertension, HbA1c level, and gender, as shown in Table 4.

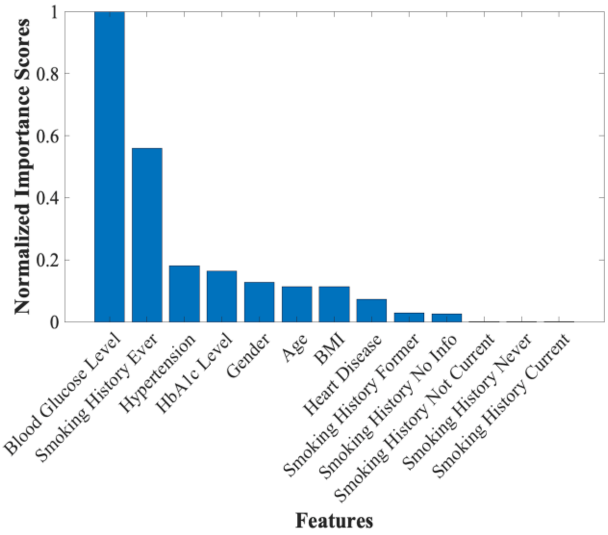

Figure 2. Normalized importance scores of individual features based on mRMR feature selection algorithm.

Table 4. Normalized importance scores of features based on mRMR feature selection algorithm.

Feature Name	Importance Scores
Blood Glucose Level	1
Smoking History Ever	0.559
Hypertension	0.180
HbA1c Level	0.163
Gender	0.127

2.3. Supervised Classifiers

The present study utilized a variety of ML-based classification models for the prediction of diabetes mellitus. The classifiers used in this study include the following: LDA, LR, Gaussian NB, Fine Gaussian SVM, Fine KNN, and EBT. A brief description of each model used for predicting diabetes mellitus is provided below.

2.3.1. Linear discriminant analysis

Linear Discriminant Analysis (LDA) is a method used for classification, which seeks to derive a linear combination of input features that effectively differentiate between distinct classes. It works by increasing the separation between the class means and reducing the within-class variance, resulting in a decision boundary that optimizes class distinction (Zhao et al., 2024).

2.3.2. Logistic regression

Logistic regression (LR) is a statistical technique employed for binary classification, which models the relationship between a dependent variable and one or more independent variables by estimating probabilities. It utilizes the logistic function to map linear combinations of the input features to values between 0 and 1, representing the likelihood of the target class. Due to its simplicity, interpretability, and effectiveness, logistic regression is widely applied in fields such as medical diagnosis and credit scoring (Anderson et al., 2003).

2.3.3. Naive Bayes

The Naive Bayes (NB) classifier is a basic probabilistic algorithm in machine learning based on Bayes' theorem. Its key feature is the assumption that the features are independent of each other when given the target class. By calculating posterior probabilities using prior class information and feature likelihoods, the algorithm efficiently classifies new instances. This approach allows it to perform well in various classification tasks, especially in high-dimensional situations where speed is important (Ren et al., 2009).

2.3.4. Support vector machine

The Support Vector Machine (SVM) methodology identifies optimal decision boundaries in multidimensional spaces by leveraging geometric principles to maximize class separation. The algorithm's fundamental idea is margin maximization, in which it finds a hyperplane that keeps the greatest distance between the closest training data points of various classes. By projecting data into higher-dimensional spaces through kernel transformations, SVM goes beyond linear classification and makes it possible to solve intricate, non-linearly separable classification problems while preserving computational efficiency (Chandra and Bedi, 2018).

2.3.5. K-nearest neighbor

The k-nearest neighbor (KNN) classifier is a straightforward and effective supervised machine learning algorithm widely applied to classification tasks. It predicts the class of a data point by analyzing the majority class among its closest neighbors, identified using distance metrics such as Euclidean or Manhattan distance (Hidayati and Hermawan, 2021).

2.3.6. Ensemble classifier

ML models known as ensemble classifiers (EC) combine several separate models to improve classification performance and accuracy. Compared to utilizing a single model, combining the strengths of several models yields better outcomes, increased reliability, and improved generalization. Boosting, stacking, and bagging are common ensemble approaches that each have a distinct function in prediction (Mohapatra et al., 2023). The hyperparameters of each implemented algorithm are presented in Table 5.

Table 5. Description of hyperparameters used for the implemented machine learning models

Supervised Classifiers	Hyperparameters	Parameter Values/description
LDA (Linear Discriminant Analysis)	Discriminant Type	Linear
	Covariance Structure	Full
	Gamma	0 (No Regularization)
	Fill Coefficients	Off
LR (Logistic Regression)	Model Type	Logistic Regression
	Distribution	Binomial
	Link Function	Logit
	Loss Function	Logarithmic Loss
NB (Gaussian Naive Bayes)	Distribution Name for Numeric Predictors	Gaussian
	Distribution Name for Categorical Predictors	mvmn (Multinomial)
SVM (Fine Gaussian Support Vector Machine)	Kernel Function	Gaussian
	Box Constraint Level	1
	Kernel Scale	0.56
	Multiclass Method	One-vs-One
	Standardize Data	No
KNN (Fine K Nearest Neighbor)	Number of Neighbors (K)	1
	Distance Metric	Euclidean
	Distance Weight	Equal
	Standardize Data	No
EBT (Ensemble Boosted Trees)	Ensemble Method	AdaBoost
	Learner Type	Decision Tree
	Maximum Number of Splits	20
	Number of Trees (Learners)	30
	Learning Rate	0.1
	Number of Predictors to Sample	Select All

3. Results

In this section, the results obtained from various ML-based classifiers are presented in detail. The quantitative assessment of classification models is conducted through the systematic analysis of performance metrics computed from the confusion matrix. The performance of those models with feature selection is evaluated using the following metrics: accuracy, sensitivity (recall or true positive rate, TPR), specificity (selectivity or true negative rate, TNR), precision (positive predictive value, PPV), negative predictive value (NPV), F1 score, and the area under the receiver operating characteristic curve (AUROC). These evaluation metrics are computed by the following equations (5-10):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

$$Sensitivity(Recall) = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity(Selectivity) = \frac{TN}{TN + FP} \tag{7}$$

$$Precision(PPV) = \frac{TP}{TP + FP} \tag{8}$$

$$NPV = \frac{TN}{TN + FN} \tag{9}$$

$$F1\ Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{10}$$

Table 6, Table 7, and Table 8 present the confusion matrices, evaluation metrics, and macro-average metrics for each classifier, respectively. The receiver operating characteristic (ROC) curves for all classifiers are illustrated in Figure 3. The macro-average evaluation metrics of each classifier are visually presented in Figure 4, with colored bar graphs showing accuracy, sensitivity, specificity, PPV, NPV, and F1 score for T2DM prediction.

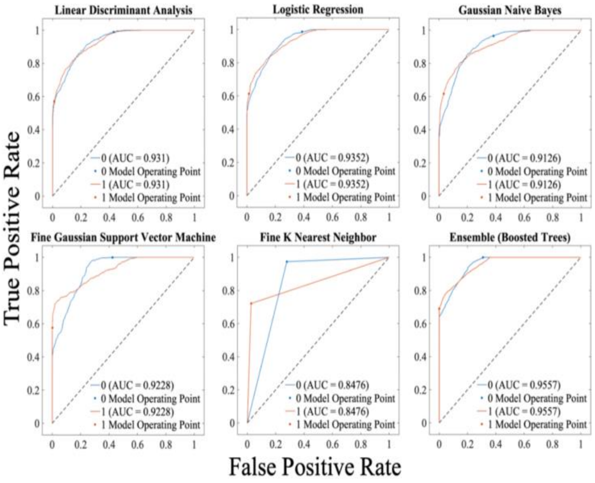

Figure 3. ROC curves for each supervised classifier with mRMR feature selection algorithm.

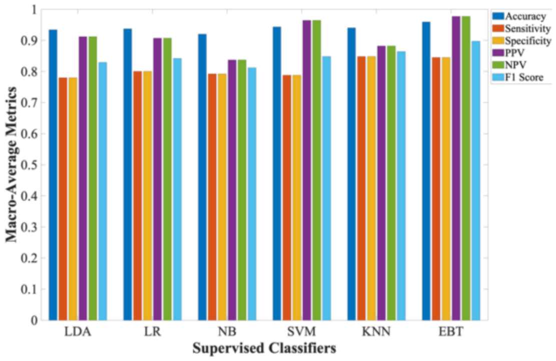

Figure 4. Macro-average evaluation metrics for each supervised classifier with mRMR feature selection algorithm

Table 6. Confusion matrices of each supervised classifier based on mRMR feature selection algorithm

Supervised Classifiers	Confusion Matrix		
	Classes	Non Diabetes	Diabetes
LDA (Linear Discriminant Analysis)	Non Diabetes	10516	118
	Diabetes	692	920
LR (Logistic Regression)	Non Diabetes	10487	147
	Diabetes	624	988
NB (Gaussian Naive Bayes)	Non Diabetes	10269	365
	Diabetes	618	994
SVM (Fine Gaussian Support Vector Machine)	Non Diabetes	10623	11
	Diabetes	683	929
KNN (Fine K Nearest Neighbor)	Non Diabetes	10354	280
	Diabetes	449	1163
EBT (Ensemble Boosted Tree)	Non Diabetes	10633	1
	Diabetes	499	1113

Table 7. Evaluation metrics of each supervised classifier based on mRMR feature selection algorithm

Supervised Classifiers	Evaluation Metrics for Each Class						
	Types	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 Score
LDA (Linear Discriminant Analysis)	Non Diabetes	0.934	0.989	0.571	0.938	0.886	0.963
	Diabetes	0.934	0.571	0.989	0.886	0.938	0.694
LR (Logistic Regression)	Non Diabetes	0.937	0.986	0.613	0.944	0.870	0.965
	Diabetes	0.937	0.613	0.986	0.870	0.944	0.719
NB (Gaussian Naive Bayes)	Non Diabetes	0.920	0.966	0.617	0.943	0.731	0.954
	Diabetes	0.920	0.617	0.966	0.731	0.943	0.669
SVM (Fine Gaussian Support Vector Machine)	Non Diabetes	0.943	0.999	0.576	0.940	0.988	0.968
	Diabetes	0.943	0.576	0.999	0.988	0.940	0.728
KNN (Fine K Nearest Neighbor)	Non Diabetes	0.940	0.974	0.721	0.958	0.806	0.966
	Diabetes	0.940	0.721	0.974	0.806	0.958	0.761
EBT (Ensemble Boosted Tree)	Non Diabetes	0.959	1	0.690	0.955	0.999	0.977
	Diabetes	0.959	0.690	1	0.999	0.955	0.817

Table 8. Macro-average evaluation metrics for each supervised classifier with mRMR feature selection algorithm (The best outcomes are highlighted in bold)

Supervised Classifiers	Macro-Average Metrics					
	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 Score
LDA (Linear Discriminant Analysis)	0.934	0.780	0.780	0.912	0.912	0.829
LR (Logistic Regression)	0.937	0.800	0.800	0.907	0.907	0.842
NB (Gaussian Naive Bayes)	0.920	0.792	0.792	0.837	0.837	0.812
SVM (Fine Gaussian Support Vector Machine)	0.943	0.788	0.788	0.964	0.964	0.848
KNN (Fine K Nearest Neighbor)	0.940	0.848	0.848	0.882	0.882	0.864
EBT (Ensemble Boosted Tree)	**0.959**	**0.845**	**0.845**	**0.977**	**0.977**	**0.897**

4. Discussion

Feature selection, while minimally impacting overall performance metrics compared to the case without feature selection (see supplementary material), plays a vital role in machine learning, especially in medical applications. It enhances model interpretability by reducing predictors, allowing healthcare professionals to better understand and trust models based on a smaller, more meaningful feature subset. This approach also improves computational efficiency, enabling faster, more scalable models for real-time decision-making. Moreover, feature selection can lead to cost savings in data gathering and processing for real-world applications. Given these benefits, this section primarily focuses on a comparative analysis of supervised classifiers with feature selection to identify the most suitable model for predicting diabetes mellitus. All models exhibited high predictive performance, with macro-average accuracy ranging from 92% to 95.9% and AUC values ranging from 84.76% to 95.57%. According to the confusion matrices in Table 5, the EBT classifier demonstrates superior performance, achieving the highest number of true negatives and the lowest number of false positives. This indicates its exceptional effectiveness in accurately identifying non-diabetic individuals while minimizing misclassification as diabetic. Although EBT exhibits a slightly higher number of false negatives compared to KNN, its overall balance between true negatives and false positives establishes it as a highly reliable model. It minimizes classification errors while maintaining a robust ability to identify diabetic cases. In contrast, KNN achieves the highest number of true positives but is limited by a substantial number of false positives, reducing its effectiveness in correctly identifying non-diabetic individuals. In addition, the results of the study reveal that the EBT classifier outperforms the other models, achieving the best accuracy of 95.9% for both non-diabetes and diabetes, an F1 score of 97.7% for non-diabetes and 81.7% for diabetes, and an AUC of 95.75% for both conditions (Table 6). Moreover, the best macro-average accuracy (95.9%), PPV (97.7%), NPV (97.7%), and F1 score (0.897) were achieved with the EBT classifier, indicating superior predictive performance compared to the other classifiers (Table 7). Additionally, the highest macro-average sensitivity and specificity were obtained, with 84.5% for the EBT classifier and 84.8% for the fine KNN classifier. The predictions produced by the ML models in this study exhibit a satisfactory level of accuracy in comparison to the results from prior research. For instance, Rastogi and Bansal (2023) utilized several data mining techniques for the prediction of diabetes based on an open-source dataset. In their study, they obtained the best accuracy of 82.46% using LR and the sensitivity of 68.88% using RF. These values are lower than those presented in Table 6 in this study. In addition, Tasin et al. (2022) performed a comparative analysis of various ML models for predicting diabetes using an open-source dataset. In their study, they employed SMOTE and ADASYN preprocessing techniques to address the problem of class imbalance. Their findings indicated that the XGBoost classifier achieved an accuracy of 81% and F1 score and AUC values of 0.81 and 0.84, respectively, which represent the best outcomes in their study but are lower than the values obtained in this study. Similarly, Tigga and Garg (2020) compared to performance of a variety of ML models for the prediction of T2DM. According to their experimental findings, the RF model has the highest accuracy of all the models, at 94.10%. Their result is also lower than the results obtained in this study. Unlike these studies, which employed resampling techniques to balance class distribution, this approach preserved the natural class imbalance to ensure real-world applicability. The observed imbalance reflects the actual prevalence of Type 2 Diabetes Mellitus (T2DM) in the population being studied, and by maintaining this distribution, the model's performance metrics more accurately represent what would be expected in real-world clinical applications. Modifying the class proportions could introduce bias and reduce the generalizability of the model to clinical practice. Furthermore, a key objective of this research was to evaluate the inherent robustness of various supervised classifiers when faced with naturally occurring class imbalance. To mitigate potential biases introduced by the imbalance, we focused on performance metrics that are less sensitive to skewed class distributions, such as the F1-score, AUC-ROC, precision, and recall. Despite the class imbalance, our evaluation metrics—including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-score—indicated that the models performed reliably across different classifiers. The results of the analysis indicate that ML-based classifiers, particularly the EBT classifier, possess enhanced predictive capabilities for estimating T2DM. The study suggests that ML-based classifiers offer a reliable approach for accurately predicting T2DM and supporting medical professionals in clinical decision-making.

5. Conclusion

This study provides a comprehensive analysis of diabetes mellitus detection, particularly T2DM, based on a large, open-source dataset obtained from Kaggle. To my knowledge, this is the first study conducted using this dataset, specifically focusing on patients aged 35 years or older for T2DM prediction. The dataset was preprocessed as an initial step to prepare it in a suitable format for ML models. In addition, a feature selection method was applied to the preprocessed dataset to identify key attributes, which were used as input for the ML models. The mRMR feature selection method was chosen due to its advantages, including balancing relevance and redundancy, applicability to different data types, robustness against overfitting, and ability to handle high-dimensional datasets. Five significant features, namely blood glucose level, smoking history ever, hypertension, HbA1c level, and gender were selected using the mRMR feature selection method. Various supervised classifiers were employed to distinguish diabetes mellitus from non-diabetes mellitus. The performance of these models was assessed using multiple evaluation metrics, including accuracy, sensitivity, specificity, PPV, NPV, F1 score, and AUROC. The findings reveal that the EBT classifier surpasses the other models, recording the highest macro-average values for accuracy (95.9%), PPV (97.7%), NPV (97.7%), and F1 score (0.897), along with the superior AUC of 95.57% for both diabetes and non-diabetes cases. Although the outcomes of the ML classifiers employed in this study provide optimal results, there remains potential for improvement in sensitivity, specificity, and F1 score. This can be achieved by further optimizing the ML models. Additionally, a more comprehensive dataset may be utilized to increase the predictive power of the models. Furthermore, future research will also include the development of hybrid ML model(s) to enhance predictive performance by integrating the strengths of various algorithms, thereby improving accuracy and robustness in predicting outcomes related to diabetes mellitus.

Author Contributions

The percentages of the author' contributions are presented below. The author reviewed and approved the final version of the manuscript.

	O.K.
C	100
D	100
S	100
DCP	100
DAI	100
L	100
W	100
CR	100
SR	100
PM	100
FA	100

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

Conflict of Interest

The author declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans. The study employed an open-source dataset.

References

American Diabetes Association. 2021. URL: https://diabetes.org/newsroom/latest-ada-annual-standards-of-care-includes-changes-to-diabetes-screening-first-line-therapy-pregnancy-technology (accessed date: January 6, 2025)

Anderson RP, Jin R, Grunkemeier GL. 2003. Understanding logistic regression analysis in clinical reports: an introduction. Ann Thorac Surg, 75(3): 753–757.

Bhat SS, Banu M, Ansari GA, Selvam V. 2023. A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms. Healthc Anal, 4: 100273.

Borse SP, Chhipa AS, Sharma V, Singh DP, Nivsarkar M 2021. Management of type 2 diabetes: current strategies, unfocussed aspects, challenges, and alternatives. Med Princ Pract, 30(2): 109–121.

Cano-Cano F, Gómez-Jaramillo L, Ramos-García P, Arroba AI, Aguilar-Diosdado M. 2022. IL-1β implications in type 1 diabetes mellitus progression: systematic review and meta-analysis. J Clin Med, 11(5): 1303.

Carrillo-Larco RM, Guzman-Vilca WC, Xu X, Bernabe-Ortiz A. 2024. Mean age and body mass index at type 2 diabetes diagnosis: pooled analysis of 56 health surveys across income groups and world regions. Diabet Med, 41(2): e15174.

Chandra MA, Bedi SS. 2018. Survey on SVM and their application in image classification. Int J Inf Technol, 13(5): 1–11.

Costa-Cordella S, Luyten P, Cohen D, Mena F, Fonagy P. 2021. Mentalizing in mothers and children with type 1 diabetes. Dev Psychopathol, 33(1): 216–225.

Fazakis N, Kocsis O, Dritsas E, Alexiou S, Fakotakis N, Moustakas K. 2021. Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access, 9: 103737–103757.

Febrian ME, Ferdinan FX, Sendani GP, Suryanigrum KM, Yunanda R. 2023. Diabetes prediction using supervised machine learning. Procedia Comput Sci, 216: 21–30.

Hidayati N, Hermawan A. 2021. K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation. J Eng Appl Sci Technol, 2(2): 86–91.

International Diabetes Federation. 2021. URL: https://diabetesatlas.org/atlas/tenth-edition/ (accessed date: January 3, 2025)

Janiesch C, Zschech P, Heinrich K. 2021. Machine learning and deep learning. Electron Markets, 31(3): 685–695.

Jordan MI, Mitchell TM. 2015. Machine learning: trends, perspectives, and prospects. Science, 349(6245): 255–260.

Kaggle. 2024. Diabetes prediction dataset. URL: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset (accessed date: October 17, 2024)

Kurt O. 2024. Model-based prediction of water levels for the Great Lakes: a comparative analysis. Earth Sci Inform, 17(3): 3333–3349.

Laakso M, Kuusisto J. 2014. Insulin resistance and hyperglycaemia in cardiovascular disease development. Nat Rev Endocrinol, 10(5): 293–302.

Lin X, Xu Y, Pan X, Xu J, Ding Y, Sun X, Song X, Ren Y, Shan PF. 2020. Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. Sci Rep, 10(1): 14790.

Ma CX, Ma XN, Guan CH, Li YD, Mauricio D, Fu SB. 2022. Cardiovascular disease in type 2 diabetes mellitus: progress toward personalized management. Cardiovasc Diabetol, 21(1): 74.

Modak SKS, Jha VK. 2024. Diabetes prediction model using machine learning techniques. Multimed Tools Appl, 83(13): 38523–38549.

Mohapatra SK, Das A, Mohanty MN. 2023. Application of ensemble learning–based classifiers for genetic expression data classification. Data Science for Genomics, Academic Press, pp: 11–23.

Rastogi R, Bansal M. 2023. Diabetes prediction model using data mining techniques. Meas Sens, 25(21): 100605.

Ren J, Lee SD, Chen X, Kao B, Cheng R, Cheung D. 2009. Naive Bayes classification of uncertain data. Ninth IEEE Int Conf Data Min, pp: 944–949.

Talukder MA, Islam MM, Uddin A, Kazi M, Khalid M, Akhter A, Moni MA. 2024. Toward reliable diabetes prediction: innovations in data engineering and machine learning applications. Digit Health, 10: 1–26.

Tasin I, Nabil TU, Islam S, Khan R. 2022. Diabetes prediction using machine learning and explainable AI techniques. Healthc Technol Lett, 10(1–2): 1–10.

Tigga NP, Garg S. 2020. Prediction of type 2 diabetes using machine learning classification methods. Procedia Comput Sci, 167: 706–716.

Varma KVSRP, Rao AA, Lakshmi TSM, Rao PVN. 2014. A computational intelligence approach for a better diagnosis of diabetic patients. Comput Electr Eng, 40(5): 1758–1765.

WHO. 2024. World health organization. URL: https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed date: January 3, 2025)

Zhao S, Zhang B, Yang J, Zhou J, Xu Y. 2024. Linear discriminant analysis. Nat Rev Methods Primers, 4(1): 70.

Zhou H, Xin Y, Li S. 2023. A diabetes prediction model based on Boruta feature selection and ensemble learning. BMC Bioinform, 24(1): 224.