Enhancing Diabetes Prediction: An Improved Boosting Algorithm for Diabetes Prediction

Md. Shahin Alam, Most. Jannatul Ferdous, Nishat Sarkar Neera
Department of Computer Science and Engineering
Bangladesh University of Business and Technology (BUBT),
Rupnagar, Mirpur-2, Dhaka-1216, Bangladesh

***Abstract*—Diabetes is increasing gradually due to the inability to effectively use the human body's insulin, which threatens public health. People with diabetes who go undiagnosed at early stages or who have diabetes have a high risk of heart disease, kidney disease, eye problems, stroke, and nerve damage for which diabetes diagnosis is crucial to prevent. Our advanced machine learning algorithm is the gateway to a revolutionary possibility of detecting whether the human body has diabetes. Developed this method based on machine learning with one lakh data and the main objective of creating a new and novel diabetes prediction model named moderated Ada-Boost(AB) that can accurately diagnose diabetes. About 10 different classification methods are applied in this research such as Random forest classifier (RF), logistic regression (LR), decision tree classifier (DT), support vector machine (SVM), Bayesian Classifier (BC) or Naive Bayes Classifier (NB), Bagging Classifier (BG), Stacking Classifier (ST), Moderated Ada-Boost(AB) Classifier, K Neighbors Classifier (KN) and Artificial Neural Network (ANN). The crucial contribution is to find out the appropriate values for the different models using the hyper-parameter tuning process. We have proposed a new boosting model named Moderated Ada-Boost(AB) which is the combination of the hyper-parameter tuned random forest model and Ada-boost model. Different evaluation metrics such as accuracy, precision, recall, f1 score, and others are used to evaluate the performance of the models. Our proposed new boosting algorithm named Moderated Ada-Boost(AB) provides better accuracy than other models whose training accuracy is 99.95% and testing accuracy is 98.14%.**

***Keywords*—Diabetes prediction; ensemble technique; machine learning; binary classification; Moderated-AdaBoost;**

I. Introduction

Diabetes is a disease that causes many diseases in the human body, resulting in reduced life expectancy and premature death due to which the death rate is increasing day by day. One of the main causes of diabetes in the human body is insulin deficiency. The foods that humans consume to sustain life inhibit the production of energy from food sources when insulin is deficient. When the human body cannot produce enough insulin or use it properly or both. It is a major cause of diabetes in the human body. When the human body develops diabetes, it is no longer possible to remove it. As a result, millions of people worldwide are going through a difficult time. As their physical condition deteriorates, they have to change their diet and exercise excessively. When the amount of sugar in their body increases, the level of diabetes in their body becomes too high, so it is no longer possible to eliminate diabetes from the body for life. 537 million people worldwide had diabetes in 2021, of whom 81% lived in low- and middle-income countries. Diabetes-related deaths totaled 6.7 million, and the cost of diabetes-related medical bills was estimated to be USD 760 billion in 2019 and would rise to USD 845 billion by 2045 [1], [2], [3], [4]. According to IDF estimates, there are 7.1 million diabetics in Bangladesh and almost the same number of undiagnosed cases; by 2025, this number is expected to quadruple. Furthermore, in low- and middle-income nations, the cost of diabetes places a heavy weight on natural expenditures [5]. So to overcome all these problems we have developed a great method through which a person can easily check if he has diabetes or not and then take the necessary steps to cure it.

The main goal of our research is to diagnose diabetes in humans. Most people can prevent having diabetes, but once it manifests in the body, it is rarely curable. The risk of having diabetes can be decreased by early identification and lifestyle modifications. When treating a patient one-on-one, doctors can correctly determine the patient's risk of diabetes. However, screening thousands of patients with high-risk conditions presents substantial challenges for doctors. In this case, population diabetes screening requires analytical techniques. Methods involving machine learning are adaptable and can be used to address a variety of issues in a range of fields. They keep proving their adeptness in any kind of decision-making, including data analysis and pattern identification. Machine learning methods can assist in solving a few common difficulties among the multitude of challenges that exist in our world. They consist of: Natural Language Processing (NLP), Optimization, Classification, Regression, Recommendation Systems, Anomaly Detection, Clustering, Language Translation, Image and Video Analysis, Time Series Forecasting, Reinforcement Learning, Healthcare, Quality Control and Anomaly Detection, Fraud Detection, Customer Churn Prediction, Content Generation, Environmental Monitoring, Personalization, Social Media Analysis, Automated Game Playing. These are only a handful of the thousands of issues that machine learning can handle; its capabilities are constantly growing and getting more sophisticated.

A. Research Contribution

This study examined a wide range of diabetes-related human health studies. Numerous research have examined the existence of diabetes in the human body. An analysis of how the human body detects diabetes or not has been attempted. The contributions noted below might be deemed noteworthy:

- To find out the best parameters for different models using the hyper-parameter tuning.

- A new and novel boosting model named Moderated Ada-Boost(AB) is developed for the automatic prediction of diabetes from the structured data.
- Different performance evaluation metrics have been used to validate the performance of our proposed model named Moderated Ada-Boost(AB).

B. Organization of this Paper

The remainder of the document is structured as follows: In Section II, the literature review was covered. In Section III, the methodology—which includes our suggested model—has been succinctly outlined. Section IV contains an analysis of the outcome. Section V concludes with a remark on future work and conclusions.

II. Literature Review

Healthcare researchers have used a variety of approaches, such as machine learning and data mining, to evaluate different datasets to predict diabetes. Notable methods include classification techniques like Naïve Bayes and Decision Trees, hybrid models that include clustering and classification algorithms like C4.5 decision trees, Neural Networks, and Random Forest Classifier, and Hadoop and MapReduce for economical analysis [6]. Random Forest (RF) surpassed Support Vector Machine (SVM) and deep learning (DL) in the comparison evaluation of machine learning and deep learning algorithms for diabetes prediction, obtaining the greatest overall accuracy of 83.67% in diabetic categorization. SVM achieved a prediction accuracy of 65.38% [7]. The paper builds a prediction model using three different algorithms. which are random forest, support vector machine, and logistic regression. With an accuracy of about 84%, Random Forest is the best algorithm for predicting Diabetes [8]. Priyanka Sonar and Prof. K. JayaMalini has presented algorithms like SVM, and ANN for identifying diabetes using ML algorithm [9]. Through the study of diabetes patient databases, researchers looked into the use of a variety of machine learning algorithms, like Random Forest, ensemble supervised learning, SVM, Logistic regression, ANN, Bayesian, and KNN, for the prediction of diabetes. We can observe from this study that the random forest classifier works more effectively than the others [10].

The goal of this study is to predict diabetes utilizing a variety of data mining classification techniques, such as KNN, Decision Trees, and Naive Bayes. The focus of the study is on predicting diabetes with high accuracy and maybe saving lives. A variety of algorithms are used for medical data for early identification [11]. Using the Pima Indian Diabetes dataset (PIDD), several researchers have used a variety of machine learning techniques, including artificial neural networks (ANN), bootstrap aggregating, adaptive boosting, decision trees, logistic regression, Naive Bayes, and Random Forest, to predict diabetes. The findings show accuracies between 75.7% and 77.21%. Various research highlights the importance of different aspects and uses different feature reduction approaches to get the optimal predictions [12]. Machine learning techniques like Adaboost, Bagging, Decision Tree, Genetic Programming, Artificial Neural Network, and Random Forest are used in several studies (Sajida, Orabi, Pradhan, Rashid, and Nongyao, among others) to predict diabetes. The results indicate that Adaboost performs better than Bagging and Decision Tree, Decision Tree, and Genetic Programming provide satisfactory results with high accuracy, and Random Forest is the most efficient algorithm among the ones used [13]. In a 4-node Hadoop cluster setting, the random forest method provides the greatest accuracy at 94% compared to the decision tree and naïve bayes algorithms [14]. Data mining techniques, particularly when combined with machine learning, have demonstrated superior predictive capabilities, accuracy, and precision when compared to traditional methodologies, as evidenced by previous studies emphasizing their effectiveness, particularly in the context of driving prediction models for conditions such as diabetes [15]. Diabetes classification methods utilized include Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbour (KNN), and Naive Bayes. Naïve Bayes, SVM, and Decision Tree classifiers are used to predict diabetes mellitus; of these, Naïve Bayes is the most effective with an accuracy of 76.3%; K-Nearest Neighbour and Logistic Regression classifiers with Gradient Boosting feature selection are also used for diabetes prediction [16].

Numerous studies have been conducted in the literature using various datasets and methods for the identification of diabetes. For example, Zou et al. used a dataset from Luzhou, China, applied PCA and mRMR for dimensionality reduction, and showed that an RF classifier achieved the highest accuracy of 80.84%. Maniruzzaman et al. used the Pima Indian diabetes dataset, applied a variety of classifiers, and discovered that an RF-based classifier with feature selection achieved the highest accuracy of 92.26%. Furthermore, Ahuja et al. employed LDA for feature selection using the Pima Indian diabetes dataset, and they found that the best accuracy of 78.70% was obtained when an LDA was combined with a Multi-Layer Perceptron (MLP) classifier. Without using feature selection, Sisodia et al. used SVM, Naive Bayes (NB), and DT classifiers and obtained the maximum accuracy of 76.30% [17]. For diabetes prognosis, the suggested approach used a unique type o[f deep neural network to boost prediction accuracy. Using the PID Data Set, the experiment revealed that the suggested approach had an accuracy of 88.41% [18]. To predict GDM in model A, the fundamental feature set was utilized, which included the patient's age, heart rate, blood pressure, and other vital indicators. The performance of EPM is satisfactory (Accuracy = 0.902%, AUC = 0.912%). With the addition of weight and gestenail changes, Model B utilized the same feature (Accuracy = 0.957%, AUC = 0.942%) [19].

Diabetes is a rapidly spreading disease with serious consequences such as cardiovascular disease and renal failure. Early diagnosis is crucial but challenging due to limited labeled data and unreliable clinical datasets. To address this, a diabetes dataset from Bangladesh has been provided along with a weighted ensemble of machine-learning classifiers. Hyper-parameter optimization and feature selection techniques are utilized to improve prediction accuracy. The proposed ensemble model (DT + RF + XGB + LGB) combined with statistical imputation and RF-based feature selection yielded the best results for early diabetes prediction. The dataset will contribute to the development of reliable machine-learning models for diabetes prediction using population-level data [20]. Diabetes is a chronic illness that is on the rise and may be quite dangerous if not caught in time. By establishing automated

methods for diagnosing diabetes patients, recent developments in machine learning techniques and ontology-based approaches have made a significant contribution to the area of medical science. Decision Tree, Naive Bayes, KNN, SVM, and ANN are among the most widely used techniques that are compared and reviewed in this study. The outcomes are assessed using performance metrics like as F-measure, recall, accuracy, and precision. According to this study's findings, SVM attains the maximum accuracy [diabetes prediction using machine learning] [21].

Since diabetes has an impact on everyone's health, it is a major worldwide problem. Using big data analytics and machine learning, researchers have been working to create an effective diabetes prediction model. Based on their research, an intelligent framework for diabetes prediction is proposed in this article. For diabetes prediction, the authors assess support vector and random forest machine learning models based on decision trees. Health professionals, stakeholders, students, and researchers interested in diabetes prediction research and development may all benefit from their creation of a novel intelligent diabetes mellitus prediction framework (IDMPF). With a minimal mistake rate, the suggested effort achieves 83% accuracy [22].

Diabetes mellitus is a metabolic disease marked by elevated blood glucose levels as a result of the body's failure to produce or react to insulin. Diabetes can cause major problems that harm essential organs if it is not addressed. Although machine learning can be used to predict diabetes, more work has to be done in this area of computational diagnosis research. Using two datasets, this research suggests a machine learning paradigm for diabetes diagnosis and prediction. Feature selection and missing value imputation techniques can be used to improve classification model accuracy. The approach uses polynomial regression and Spearman correlation for missing value imputation and feature selection, respectively. A custom deep neural network, support vector machines, random forests, and other machine learning models are proposed for classification. Grid search and cross-validation are used in the models' optimisation. The proposed deep neural network model provides good accuracy in diabetes prediction, according to experimental results on two datasets. The framework's classifiers and preprocessing techniques perform better than those of other approaches. The models' source code is accessible to the general public [23].

In this work, they employed K-NN, DT, LR, BNB, and SVM—five of the most widely used algorithms for identifying and categorizing binary issues, like diabetes. The maximum accuracy attained by the K-NN model was 79.6% [24]. Using the PID and HFD datasets, the CFA was compared to the GA. To the best of the information we have, the only metaheuristic algorithm for type 2 diabetes detection is the GA. Six classifiers were used to test the CFA and GA algorithms: K-NN, RF, DT, LR, SVM, and NB. Of these, rf and KNN provided the highest accuracy, at 77% and 79%, respectively [25]. Diabetes of either type could be detected most accurately by the machine learning models, which produced AUROC and AUPR curves of 0.84% (95% CI 0.76%, 0.91%) and 0.84% (95% CI 0.78%, 0.93%), respectively. For diabetes, the model's sensitivity and specificity were 0.82% and 0.75%, respectively. Comparable results were established for type 1 (AUROC 0.81% and AUPR 0.72%) and type 2 (AUROC 0.88% and AUPR 0.81%) diabetes, as well as $HbA1c \geq 6.5\%$ [26]. ML has drawn more and more interest in recent years from a variety of study domains. Of all the machine learning approaches available today, ANNs are performing especially well in positions related to health [27]. In this study, we describe a unique no-prop technique that uses a multi-layer neural network to classify the three forms of diabetes mellitus. A multi-layer neural network is used to improve the efficiency of categorization. The best specificity and sensitivity values of 0.95% were achieved by the suggested multi-layer neural network [28].

The objective of this study is to apply non-invasive techniques to identify diabetes and prediabetes. To do this, they used machine learning in conjunction with ECG. The study made use of clinical data from 1262 people who were part of the Diabetes in Sindhi Families in Nagpur study. Three sets of the dataset were created: training, validation, and test. After processing the ECG recordings, minority oversampling was used to balance the training dataset. Based on the processed ECG data, the classifier was trained to predict whether a person will belong to the prediabetes, type 2 diabetes, or "no diabetes" groups. The American Diabetes Association's definition of the requirements for these classes was followed [29]. According to the SHAP, glucose is the specific factor that most influences the possibility of developing diabetes; however, when combined with age and body mass index (BMI), it has a far greater effect. Furthermore, BMI and the diabetes pedigree function evaluate highly for the prediction of diabetes. For this reason, if blood glucose control is difficult, attention should be directed towards managing BMI and the diabetes pedigree function. With the guidance of SHAP, we fit the ML algorithms for diabetes prediction using a new dataset that was created from the original one. Xgboost and Adaboost outperformed other models with 94.67% accuracy and F1 scores of 95.27 and 95.95, respectively [30].

During 1995, there were approximately an estimated 135 million cases of diabetes globally; by 2025, there were expected to be at least 300 million cases. Over 1995 and 2025, the number of persons with diabetes is expected to rise by 42% (from 51 to 72 million) in advanced nations and by 170% (from 84 to 228 million) in developing nations. Diabetic is associated with a number of potentially modifiable risk factors, such as insulin resistance, obesity, physical inactivity, and nutrient elements. In population at risk, diabetes may be avoidable, although the outcomes of current clinical trials are not yet known. There are presently a number of effective and affordable therapeutic options available to lessen the burden of diabetic complications, including the use of aspirin and ACE inhibitors; early identification and treatment of retinopathy, nephropathy, and foot disease; and management of blood pressure, cholesterol, and glucose. Diabetes is a serious public health issue that is starting to propagate like wildfire. While diabetes prevention may one day be achievable, there is now a great deal of possibilities to improve the use of currently available medications to lessen all the challenges connected to diabetes. Research focused at better understanding the causes of underuse of current medicines and how to improve this might be advantageous to many nations [31].

Diabetes prediction in maximum research work has mostly

employed discrete classifiers, including Random Forest, SVM, ANN, and Naive Bayes, along with simple ensemble techniques like bagging and boosting. They seldom ever investigate sophisticated hybrid ensemble methods, though, which can lead to better results. Although some studies shed light on hyperparameter tuning, many do not explain the optimization procedure in depth, which might compromise the models' efficacy and repeatability. Most studies focus on accuracy as the main metric, frequently ignoring other important performance metrics that offer a more thorough assessment of model performance, such as precision, recall, F1-score, and AUC-ROC. The lack of attention to model generalization capabilities is a frequent problem. High training accuracy is frequently reported, but testing accuracy and the overfitting danger are not sufficiently discussed, which is crucial for using these models in real-world scenarios. Furthermore, the reliability of the models has been affected by the varied handling of class imbalance, a crucial component in medical datasets, between research. Confusion matrix-based detailed assessments are often lacking, which are crucial to comprehending the kinds of inaccuracies the models make. Certain studies employ feature selection methods such as PCA and mRMR, but they don't combine them with sophisticated ensemble approaches to enhance performance even more. Furthermore, even though complicated datasets are occasionally used, sophisticated preprocessing, feature selection, and sophisticated ensemble approaches are frequently not integrated into a single, coherent workflow. Overall, to increase the accuracy and dependability of diabetes prediction models, there is a clear need for more thorough and rigorously methodical approaches that incorporate these cutting-edge strategies. Furthermore, this article demonstrates how our suggested model, Moderated-AdaBoost (AB), performs better than alternative algorithms when compared to the resilience of Artificial Neural Networks and Random Forests.

By utilizing a moderated Ada-Boost model where the hyper-parameter tuned Random Forest is used as the base estimator, our method combines the advantages of many ensemble approaches to provide a unique and reliable diabetic prediction model. To guarantee outstanding performance, we used GridSearchCV to fine-tune the Random Forest classifier's hyperparameters. Several measures, including AUC-ROC, were included in our study to give a thorough picture of the model's capacity to manage class imbalances and produce precise predictions over a range of thresholds. Strong generalization to new data is demonstrated by our excellent testing accuracy (98.14%) and training accuracy (99.95%). Our approach placed a strong emphasis on necessary preprocessing measures such as encoding, normalization, and balancing to successfully manage imbalanced datasets while reducing bias towards the class that is most prevalent. To enhance openness, replicability, and trustworthiness, we provided thorough instructions for our data pretreatment, model training, and assessment procedures.To shed light on true positives, false positives, true negatives, and false negatives along with identifying areas in demand for enhancement we implemented a confusion matrix into our study We showed that our model was superior in terms of accuracy and generalization by comparing it with other algorithms (e.g., RF, SVM, LR, NB, and KNN). Through the integration of several preprocessing approaches, effective hyperparameter tuning, and an advanced hybrid model, our methodology provides a solid solution for diabetes prediction, despite the task's complexity.

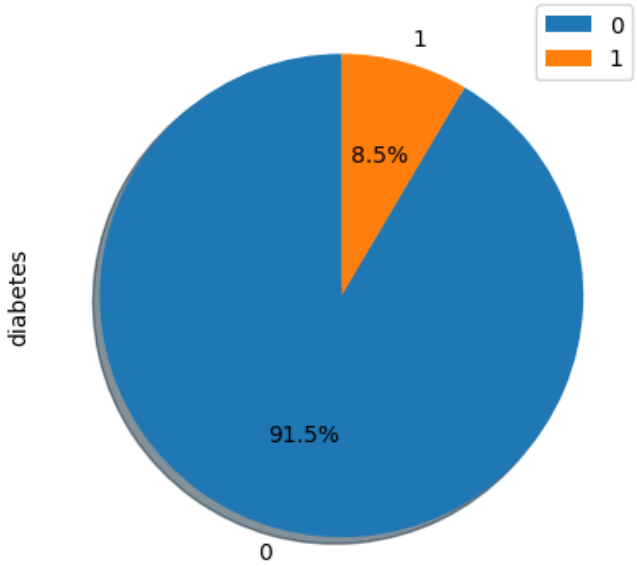

Fig. 1. Exploring how many patients are in a class.

III. Methodology

A. Datasets Description

The dataset under consideration comprises 100,000 records, of which 91,500 are non-diabetic and 8,500 are diabetic. The study made use of the "Diabetes_Prediction_Dataset" dataset. The numbers of patients with diabetes (8500) and those without (91500) are displayed in Fig. 1.

B. Dataset Preprocessing

1) Feature encoding: Standardizing categorical data into a format that works better with machine learning algorithms is accomplished with the help of this transformation which is shown in Fig. 2 and Fig. 3. The 'gender' column values in this research are converted from strings ('Female' and 'Male') to numeric values (0 and 1).

Fig. 5 similarly illustrates the conversion of data from category to numerical type.
('no info': 0 replaces 'no info' with 0,
'never': 1 substitutes 'never' with 1,
'current': 2 substitutes 'current' with 2,
'former': 3 substitutes 'former' with 3,
'ever': 4 substitutes 'ever' with 4,
'not current': 5 substitutes 'not current' with 5) in order to guarantee that the characteristics are on the same scale and that the association with the goal variable is maintained.

2) Feature scaling: Building accurate and trustworthy machine learning models demands an in-depth understanding of the distribution and relationships of the data, which can be made possible by this procedure, which provides insight into how the feature scaling process has changed the original data.

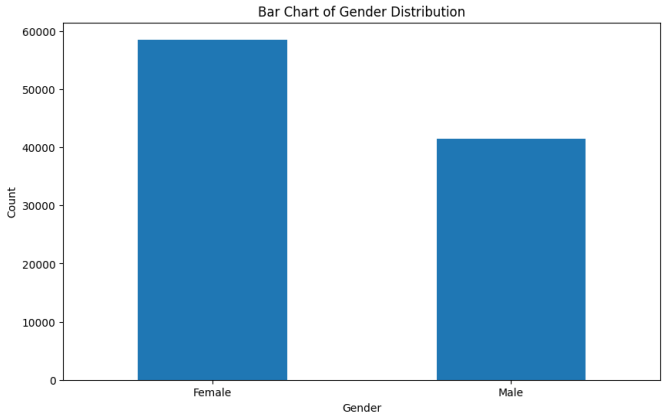

Fig. 2. Gender distributions in our datasets.

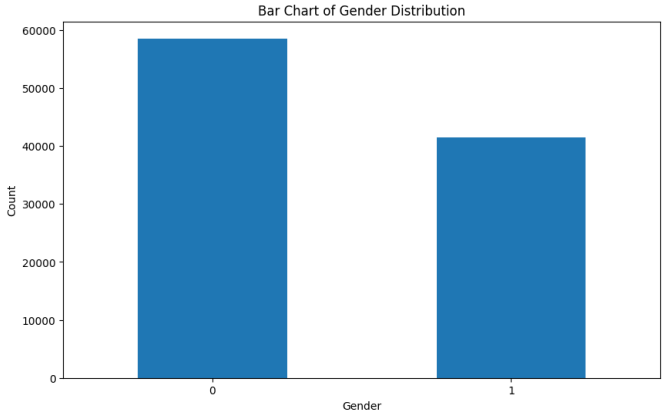

Fig. 3. Transformation of categorical to numerical type (Gender).

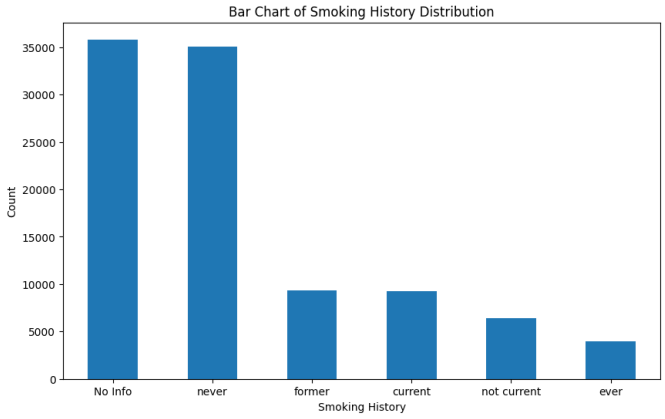

Fig. 4. Exploration of smoking history.

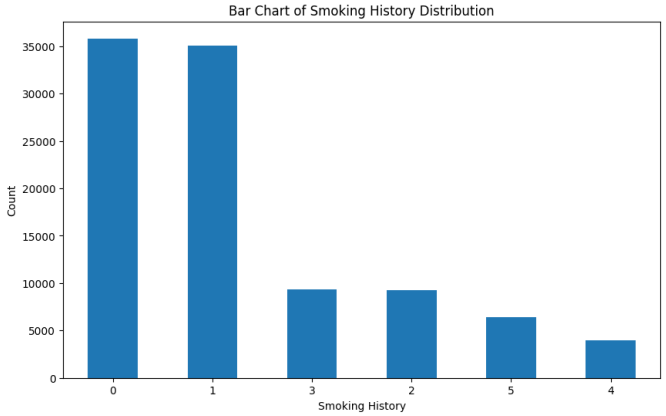

Fig. 5. Transformation of categorical to numerical type (smoking history).

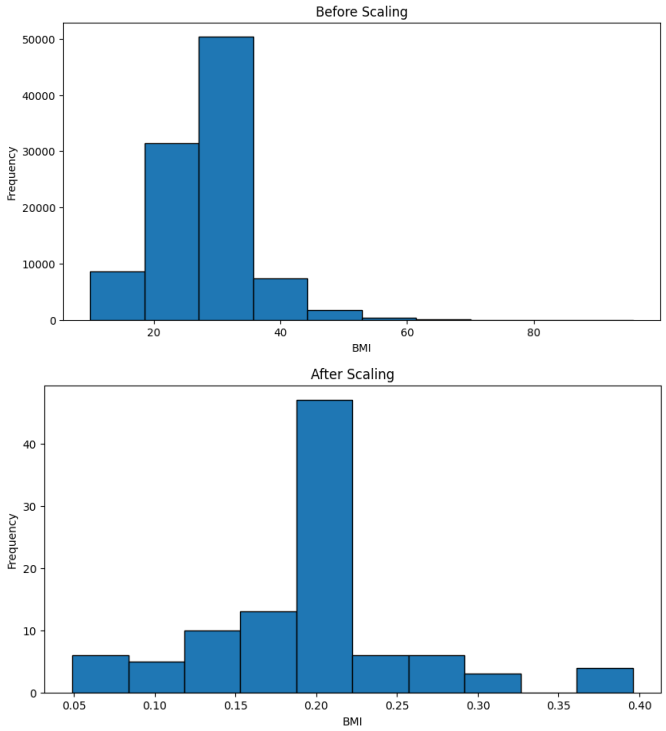

Fig. 6. Before and after feature scaling.

In general, Fig. 6 depicts a basic phase in preparing data for machine learning tasks, which guarantees that features are scaled suitably to enhance the convergence and performance of machine learning algorithms.

3) Datasets balance: Among the datasets worked on, there are 100,000 data in which the number of non-diabetic data is 91,500 and the number of data with diabetes is 8,500. Due to the imbalance of the data, the balance was done by bringing the minority class to the same level as the majority class so that the number of data with and without diabetes stood at 183000. The "diabetes_prediction_datasets" datasets were used for the investigation.

Subsequently, as seen in Fig. 7, these unbalanced datasets were balanced to equal numbers.

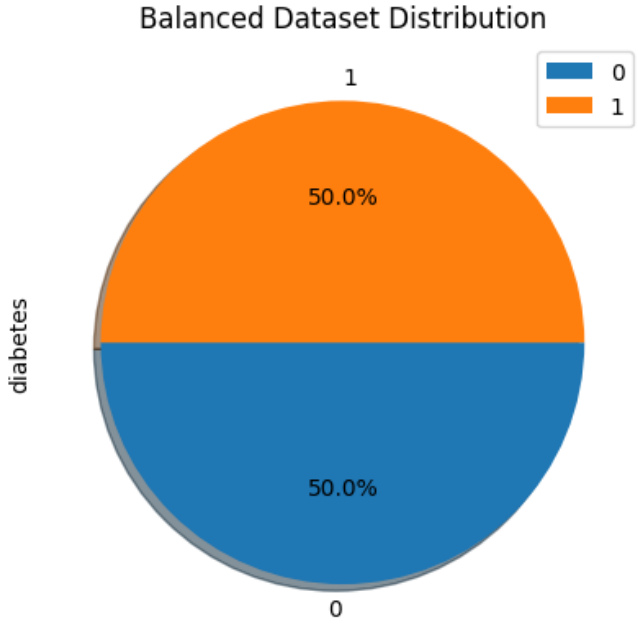

Fig. 7. Distribution the datasets between classes after balancing.

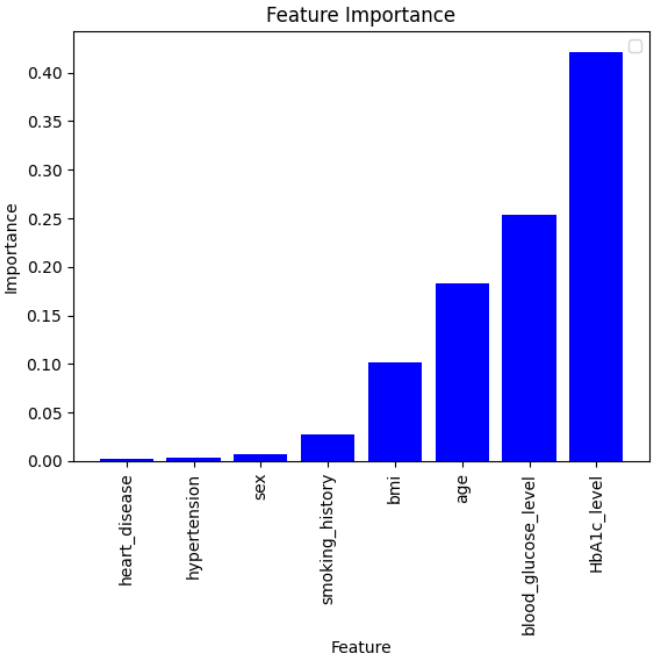

Fig. 8. Feature importance according to the datasets.

4) Feature importance: Random forest was utilized to determine the feature significance in the datasets used in this study, which had eight features. The importance of each signal in identifying or forecasting diabetes is shown in Fig. 8. This suggests that the most significant factor is the HbA1c_level, Sex, hypertension, and heart_disease are the least significant. Nevertheless, every aspect has been refined in this study.

C. All Applying Models

1) K Nearest Neighbor(KNN): The k-nearest neighbors (KNN) algorithm is a straightforward and intuitive algorithm for both classification and regression. It functions by identifying the k data points that are closest to an input data point, or its neighbors. Eq. (1) works by predicting the data based on the average value of those neighbors (for regression) or the majority of classes (for classification). The technique uses the available data to make predictions rather than explicitly training the model.

The calculation formula has been represented at:

$$\text{Predicted class for } x_{\text{new}} = \text{argmax} \sum_{i=1}^{k} I(y_i = c) \qquad (1)$$

Where:

- x_new is the new input data point.
- $k(1)$ is the number of neighbors.
- y_i is the class label of the ith neighbor.
- c iterates over all possible class labels.
- $y_i = c$ is an indicator function that evaluates to 1 if $y_i = c$, and 0 otherwise.

It's crucial to remember that the actual distance between data points and the neighbors picked are determined by the distance measure selected as well as implementation specifics. Although the KNN method is briefly described above, it's important to remember that libraries or frameworks are usually used to implement KNN since they manage computations well and offer extra capabilities for customization and optimization. The knearest neighbors (KNN) model applied to the "diabetes_prediction_dataset" dataset used to classify diabetes cases was evaluated. Based on the output, the following is an explanation of the performance of the KNN model:

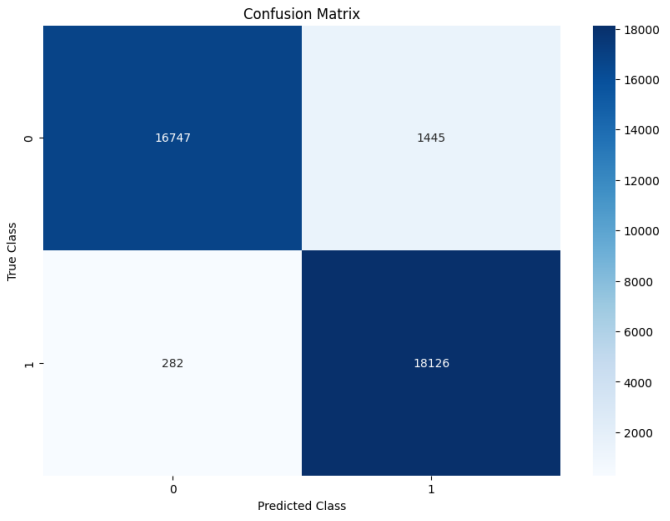

Fig. 9. Confusion matrix of KNN model.

The KNN model's overall accuracy was around 96.48%. The confusion matrix in Fig. 9 represents each example dataset

that we have acquired. this accuracy shows the percentage of properly identified occurrences. For every class, the classification report offers a more thorough analysis of the model's performance:

- Class 0: The model correctly detected instances of this class, as evidenced by its high recall (92.5) and training accuracy (0.99).
- Class 1: The model did a decent job at differentiating between instances of this class, with strong accuracy (92.5) and recall (0.99).

2) Random forest classifier: The performance of RandomForestClassifier, an extremely potent ensemble learning method, may be greatly influenced by a number of hyperparameters. In order to optimize our method, We concentrated on the following hyperparameters in the below:

- n_estimators (200): The number of trees in the forest.
- max_depth (None): Maximum depth of forest trees.
- min_samples_split (2): Minimum number of samples required to split an internal node.
- min_samples_leaf (1): Minimum number of samples required in a leaf node.

The model's accuracy throughout training was 99.95%. The test accuracy of 97.87%, however, points to a little decline in performance. Once trained, the random forest algorithm makes predictions very slowly, but it trains quickly. While a model with more trees will predict outcomes more accurately, it will also operate more slowly. We enhanced the RandomForestClassifier's capacity to identify intricate correlations in the diabetes dataset by adjusting its hyperparameters. It will be possible to contribute to a more durable and trustworthy diabetes prediction model if the chosen hyperparameters indicate a configuration that maximizes the predicted accuracy. As a crucial part of a model optimization approach, hyperparameter tuning guarantees that our machine learning model is optimized for the particular goal of diabetes prediction and produces appropriate results.

Fig. 10 demonstrates that even though 18050 positives were real positives—that is, diabetes—the model accurately predicted them to have the disease—440 individuals who did not have diabetes but were misclassified as having diabetes by the model. Additionally, 17752 is the estimated number of people without diabetes. In conclusion, 358 individuals with diabetes who had diabetes were misdiagnosed as having the disease.

3) Logistic regression hyperparameters: Numerous hyperparameters define the logistic regression method, which is frequently used for binary classification applications like diabetes prediction. We go over the hyperparameters we looked at below:

- C (1): The C parameter controls the penalty strength, which can also be effective.
- Penalty (12): The type of regularization term applied ('l1' for L1 regularization, 'l2' for L2 regularization). Note: not all solvers support all regularization terms.

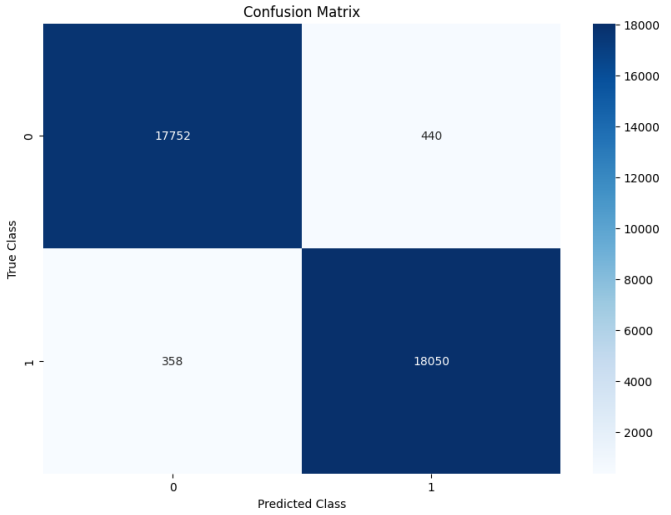

Fig. 10. Confusion matrix of random forest classifier.

- Solver (liblinear): Algorithm to use for optimization ('liblinear' is suitable for small datasets).
- Random_state (0): Seed for reproducibility.

To systematically explore the hyperparameter space and identify the optimal combination, grid search cross-validation was employed. To do this, a grid of potential hyperparameter values had to be generated, and the model's performance had to be evaluated for each combination using cross-validation. Following the grid search process, we were able to identify the optimal Logistic Regression model and the associated hyperparameter values. We carefully tweaked the Logistic Regression hyperparameters until we found the model configuration that maximized the model's accuracy in predicting our diabetes dataset. By using hyperparameters that compromise between regularization strength and model complexity, the model is ensured to be well-suited to the underlying patterns in the data.

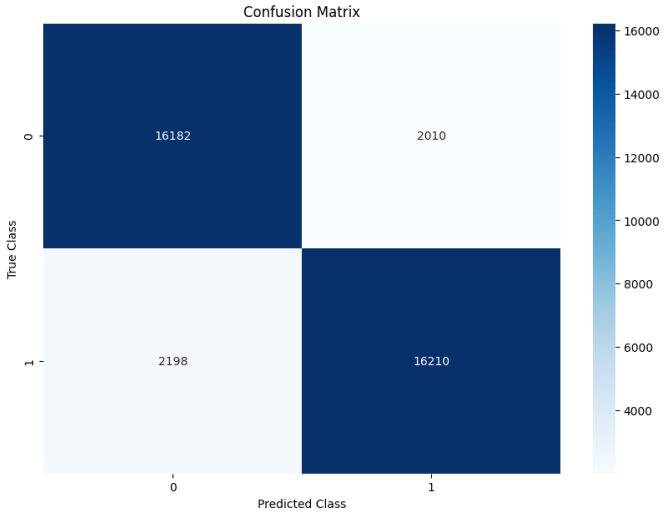

Fig. 11. Confusion matrix of logistic regression.

While 16,210 positivity were true positives, meaning that

the approach precisely determined that they had diabetes, Fig. 11 demonstrates that in 2010 people were not suffering from diabetes but were mistakenly labeled to be suffering from diabetes by the predictive algorithm. Furthermore, the expected number of individuals without diabetes is 16182. In summary, 2198 individuals with diabetes received a false diagnosis.

In the final analysis, adjusting the logistic regression hyperparameters is an essential step in our full model optimization process that enables us to produce an exceptionally accurate diabetes prediction model.

4) Decision tree classifier: Hyperparameter tuning for the decision tree classifier was accomplished using a method called GridSearch Cross-Validation (GridSearchCV). The purpose of hyperparameter tuning is to discover the ideal combination of hyperparameter values that leads to the best performance of the model. The decision tree classifier in this instance is determined by a number of hyperparameters, including the maximum depth of the tree, the minimum number of samples needed to split an internal node, the minimum number of samples that must be present, and the splitting criterion ('Gini' or 'entropy'). A leaf node. Where:

- Criterion: entropy
- Max_depth: None
- Split: 2
- Leaf: 2

Grid search is examining various combinations of these hyperparameter values in a methodical manner. The 'cv=3' parameter indicates that 3-fold cross-validation is used for the evaluation. This involves dividing the dataset into three sections and training and evaluating the model three times, with a different subset being used as the validation set each time.

The optimal model is chosen based on the highest average cross-validated score following the grid search. The average model performance over various cross-validation folds is represented by this score. Next, the optimal model and hyperparameters are printed. By automating the process of determining a decision tree model's optimal hyperparameters, this method improves the model's predictive ability on fresh, untested data.

The remainder of Fig. 12 illustrates that 490 persons had no symptoms of diabetes but were mistakenly classified as having diabetes by the prediction algorithm, even though 17717 positives were true positives, indicating the approach accurately identified that they had the condition. Moreover, 17702 persons are predicted to be free of diabetes. In conclusion, 69 diabetic individuals were given the incorrect diagnosis.

5) Stacking classifier: An ensemble model for stacking classifiers is created to enhance a dataset's classification performance. Three different base models comprise the ensemble: support vector machines (SVM), decision trees, and logistic regression. With the distinct properties that each of these models contributes, the ensemble is able to capture both linear and non-linear correlations between the data.

- SVM, decision trees, and logistic regression are the three fundamental models that are employed. While decision trees and support vector machines (SVMs) offer robust and non-linear classification skills, logistic regression offers a linear approach.
- It is implemented with a stacking classifier that mixes the base model's predictions. A RandomForest classifier, renowned for merging predictions from several decision trees, was selected as the meta-learner.
- Evaluating generalisation ability using both training and testing datasets.
- The achieved accuracy sheds light on the model's functionality and capacity to apply previously learnt patterns to fresh data.
- Stacking combines linear and non-linear techniques to maximise the potential of several models. By combining predictions, the RandomForest meta-learner seeks to mitigate the shortcomings of individual models.
- For datasets used for training and testing, accuracy is the most important performance indicator. To find out how successfully the ensemble generalises to new data, evaluation is crucial.
- It shows to be a flexible and strong ensemble model by combining logistic regression, decision trees, stacking classifiers, and SVMs with random forest meta-learners.
- Investigate different meta-learner iterations and supplementary base models for optimisation. Adjust settings and methods to improve the overall performance of the group.
- The study adds to our understanding of the effectiveness of individual models, the process of collaborative learning, and the predictive performance attained by the stacking classifier.

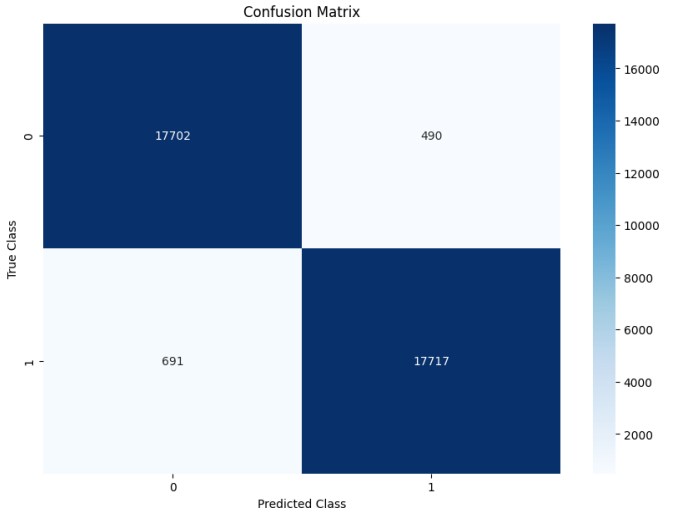

Fig. 12. Confusion matrix of decision tree.

The following section of Fig. 13 suggests that although 17754 positives were true positives, accurately recognized by the approach to be suffering from diabetes, 630 people did not

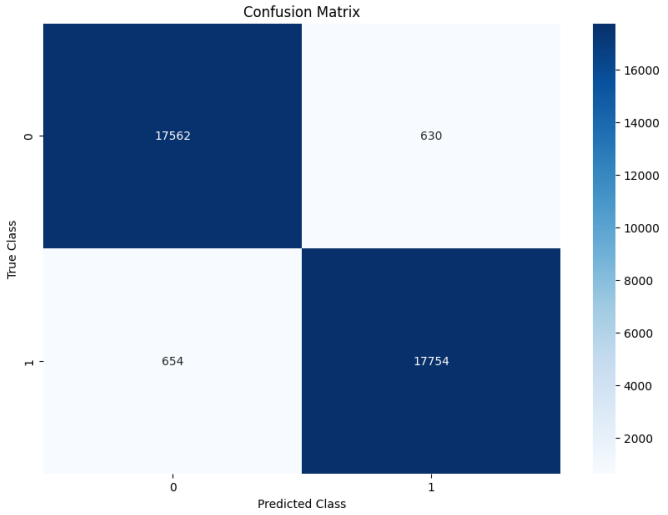

Fig. 13. Confusion matrix of stacking classifier.

exhibit any signs of the ailment, yet were mistakenly categorized as suffering from diabetes by the prediction algorithm. Moreover, it is estimated that 177562 people will be free of diabetes. To sum up, 654 people with diabetes received the incorrect diagnosis.

6) Bagging classifier: To improve predictive accuracy, an ensemble model for a bagging classifier has been created in this study. This ensemble's foundation model is a RandomForest classifier, which is well-known for its capacity to build a variety of decision trees. The RandomForest is used as the foundation model when the Bagging Classifier is first initialised, and numerous instances of the base model are generated throughout the training phase. Through bootstrap sampling, each instance is trained on a different portion of the training data, adding diversity. The Bagging technique's main advantage is its diversity, which makes the ensemble more reliable and accurate. Next, the trained Bagging Classifier is assessed using the test and training datasets. Accuracy measures are used to gauge how well the model fits the training data and how well it generalises. The outcomes, in particular the training and test accuracy, shed light on the ensemble's overall performance. With ensemble learning, this Bagging Classifier attempts to provide a better and more dependable predictive model for the provided dataset by utilising the advantages of the RandomForest model.

7) Support Vector Machine (SVM): A supervised machine learning approach that may be applied to regression and classification problems is called a support vector machine. It operates by locating the hyperplane in the feature space that best divides various classes. To improve generalization to new, untested data, the hyperplane is used to maximize the margin, or distance, between the classes.

With the application of various kernel functions, SVMs can handle both linear and non-linear separation boundaries, which makes them very useful when working with high-dimensional data. When the appropriate regularization parameter is used, they exhibit robustness against overfitting. For problems involving regression and classification, one kind of supervised machine learning technique is called Support Vector Machine (SVM).

- Accuracy: The percentage of cases in the dataset that are properly categorized out of all occurrences in the dataset. The SVM model's accuracy in this instance is around 0.90, or 90%.
- Classification Report: Each class in this classification issue is given comprehensive performance metrics in this section.

Accuracy, which measures the accuracy of positive predictions, is defined as the ratio of true positive predictions to the total predicted positives within a given class in evaluating the performance of a classification model. Conversely, recall measures the model's sensitivity to positive examples by calculating the ratio of genuine positive predictions to all real positives within a class. When dealing with unequal class distributions, the F1-Score provides a comprehensive assessment that ideally balances recall and accuracy. Regarding output, the accuracy, recall, and F1 scores for every class offer information on how well the SVM model performs for every unique class. One important finding is that a higher F1 score indicates a good balance between recall and precision, making it a useful indicator for a thorough evaluation of the model's performance.

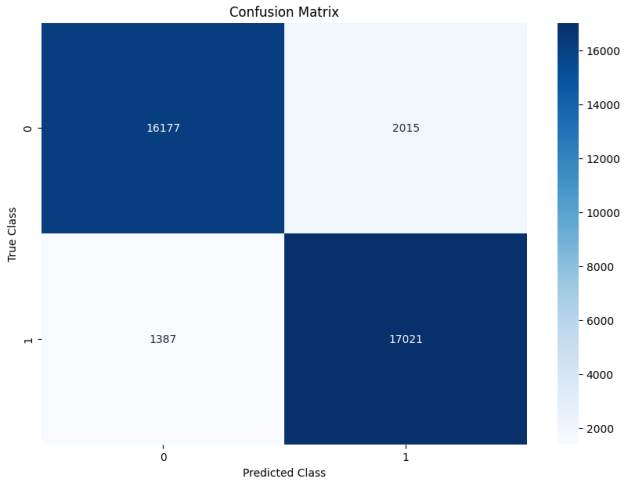

Fig. 14. Confusion Matrix of Support Vector Machine (SVM).

According to the following aspect of Fig. 14, the prediction algorithm erroneously categorized 2015 persons who did not display any signs of the disease as having diabetes, even though 17021 true positives were accurately identified as carrying diabetes by the methodology. In addition, an estimated 16177 individuals have no form of diabetes. In conclusion, 1387 diabetic patients were given the incorrect diagnosis. It has a high rate of misclassification.

8) Naive bayes classifier: For classification problems, a probabilistic machine learning technique called the Naive Bayes Classifier is employed. Based on the "naive" assumption of feature independence—that is, that all features are independent of one another given the class—it is based on the Bayes theorem. Naive Bayes classifiers frequently work well

in reality and are particularly helpful for text classification applications, despite this oversimplifying assumption. The Naive Bayes Classifier formula may be written as follows:
Eq. (2) is given by:

$$P(y|x_1,x_2,\ldots,x_n) = \frac{P(y)\cdot P(x_1|y)\cdot P(x_2|y)\cdot \ldots \cdot P(x_n|y)}{P(x_1)\cdot P(x_2)\cdot \ldots \cdot P(x_n)} \quad (2)$$

Where:

- $P(y|x_1,x_2,\ldots,x_n)$ is the posterior probability of class y given the features $x_1,x_2,\ldots,x_n$
- $P(y)$ is the prior probability of class y
- $P(xi|y)$ is the likelihood of feature x_i given class
- $(P(x_1),P(x_2),...P(x_n))$ are the marginal probabilities of the features.

In actuality, since the denominator $P(x_1)\cdot P(x_2)\cdot \ldots \cdot P(x_n)$ is constant for all classes, it may be disregarded when comparing probabilities for various classes. The instance is assigned to the class with the highest probability via the Naive Bayes Classifier, which determines the likelihood of each class given the characteristics. Naive Bayes classifiers come in a variety of forms, each having a unique method for representing the likelihoodP $(x_i|y)$, including: Gaussian Naive Bayes: Made the assumption that feature continuous values had a Gaussian distribution. Multinomial Naive Bayes: Often used for text classification where features are word frequencies, this algorithm works well with discrete data. Bernoulli Naive Bayes: For binary data, this method is comparable to Multinomial Naive Bayes. In spite of its straightforward premise, Naive Bayes is surprisingly successful, particularly when used for tasks like text categorization and other comparable ones where its efficacy and efficiency make it a popular option. The evaluation outcomes of a Naive Bayes classifier used to solve a classification issue are shown in Fig. 15. The accuracy of the model, which is around 0.835% or 83.5%, shows that it can be improved overall.

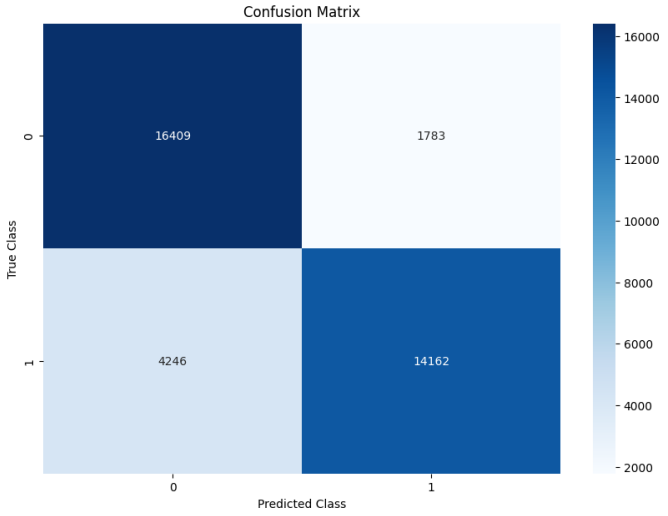

Fig. 15. Confusion matrix of naive bayes classifier.

The prediction system misclassified 1783 people who did not show any signs of the condition as having diabetes, even though 14162 genuine positives were accurately detected as suffering from the disease, which is demonstrated by one of the following characteristics of Fig. 15. Furthermore, there are about 16409 individuals who do not have diabetes. In summary, 4246 diabetic patients received the incorrect diagnosis. Its high erroneous rate is unacceptable at all.

D. Proposed Model

Hyper-parameter tuning of Random Forest Classifier using GridSearchCV. It selects and rates the best Random Forest Classifier model. Next, the Moderated Ada-Boost(AB) Classifier is constructed using the best Random Forest Classifier as its base estimator. The processing of the training and testing datasets is then shown in Fig. 16, where the Moderated-AdaBoost(AB) Classifier is trained and evaluated.

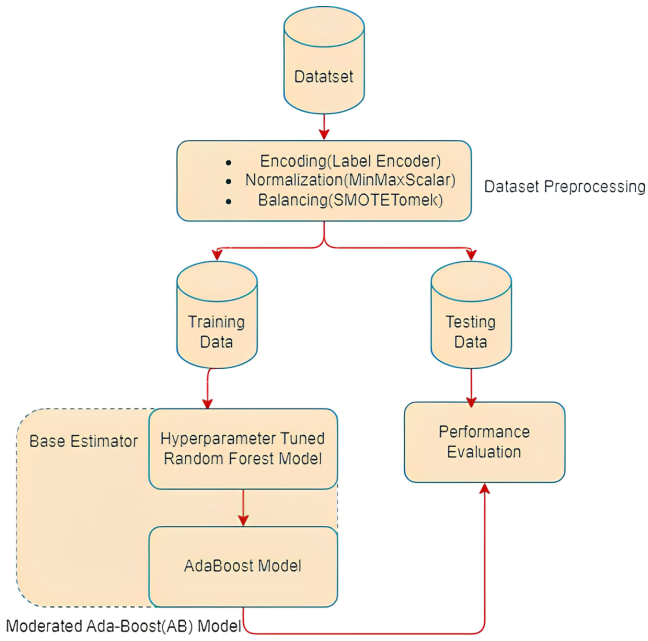

Fig. 16. The Proposed Model (Moderated Ada-Boost(AB)).

In Fig. 16, after preprocessing the dataset through some preprocessing techniques such as encoding. normalization, and balancing, the training data are used to train the proposed moderated Ada-Boost(AB) model. Then testing data is used to evaluate the performance of the model. After hyperparameter optimization yields the ideal hyperparameters for the random forest model, the random forest is chosen as the base estimator. These hyper-parameters value for the Random Forest model include:

- There are 200 trees (n_estimators).
- 'gini' is the prerequisite for splitting.
- Trees can grow to any depth (max_depth): 'None' until all their leaves are pure.
- The following factors are considered while determining the optimal split (max features): 'sqrt' (the square root of the overall number of features).
- Random state: 0 (to ensure repeatability)

Random Forest is an effective and adaptable ensemble learning technique that can minimize over-fitting and handle complicated datasets. The most recent estimator of the stacking model is the trained random forest model; the predictions made by the random forest model are input features for the Proposed Moderated Ada-Boost(AB) model, which incorporates the estimator (the beforehand trained random forest model) and the ideal parameters. Using training as well as testing datasets, the suggested moderated Ada-Boost(AB) model's performance is assessed; the results show a 99.95% training accuracy and a 98.14% testing accuracy. As a result, our proposed model named the moderated Ada-Boost(AB) model shows high test accuracy and good generalization to unseen data signifying effective power usage.

- True positive (TP): The model properly predicted these cases as positive, i.e., having diabetes, even though they were truly positive, i.e., having diabetes.
- False positives (FP): People who aren't suffering from diabetes but were mistakenly predicted by the model to have it.
- True negative (TN): This is an innovative case where the model accurately predicted the patient's absence of diabetes, even if the patient didn't suffer from the disease.
- False negative (FN): Because the model was unable to accurately forecast, it only predicted individuals who had diabetes and showed certain shortcomings, among them people who weren't diagnosed with diabetes.

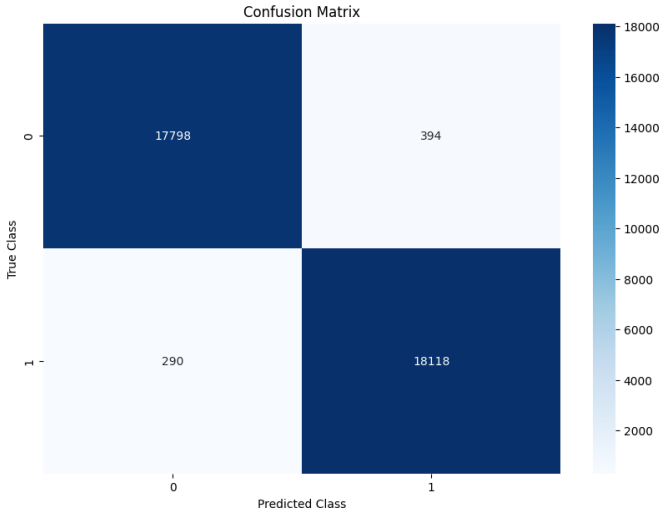

Fig. 17. Confusion Matrix of Proposed Model(Moderated Ada-Boost(AB)).

With a larger percentage of true positives and true negatives than false positives and false negatives, Fig. 17 illustrates how effectively the model works, especially in accurately recognizing positive and negative situations. It offers data on the model's performance when correctly categorized.

IV. Result and Discussion

A. Accuracy Rate of Different Algorithms

The wide range of accuracy outcomes produced by various algorithms provides a clear picture of each one's advantages and disadvantages. Our proposed model the Moderated Ada-Boost (AB) is a very effective front-end performer; Fig. 18 shows an astounding 99.95% accuracy for the training phase. The capacity of the model to identify intricate patterns and characteristics within datasets is supported by this consistency of accuracy. This widely held disagreement highlights the algorithm's capacity to understand intricate linkages while preserving good generalization.

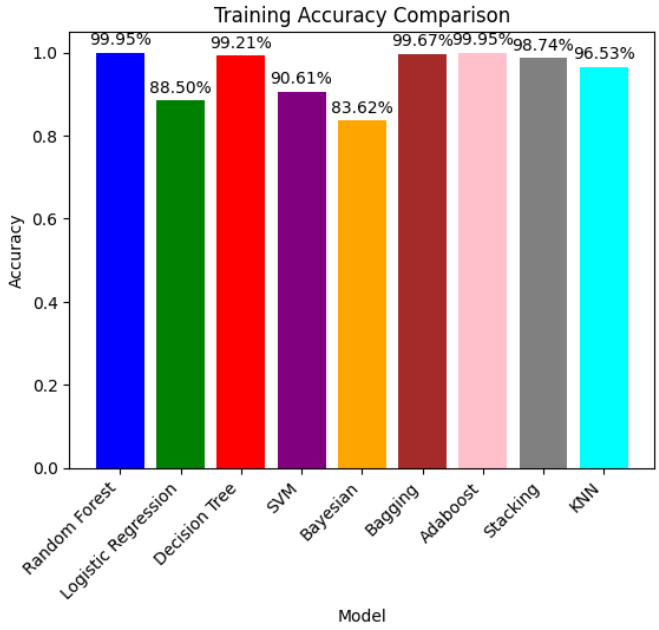

Fig. 18. Comparative training accuracy across models.

While bagging offers the highest certainty, Decision Tree provides a pretty close accuracy in training, and Random Forest and Moderated Ada-boost(AB) have the same assurance. However, in the case of Moderated Ada-boost(AB), shown in Fig. 19, we obtained the maximum accuracy throughout the testing which is 98.14%. This leads us to the conclusion that our suggested moderated Ada-Boost(AB) provides the highest level of trust and the best backing.

B. Confusion Matrix

A thorough summary of the classifier's performance for every class is given by the confusion matrix. This illustrates its advantages and disadvantages in terms of identifying individuals with and without diabetes. This matrix is a useful starting point for computing several performance measures, including accuracy, precision, recall, and F1-score for every class, giving information about the classifier's overall performance as well as potential areas for improvement.

Recall, accuracy, and F1 score are three metrics that were used in this study to evaluate the model's performance in

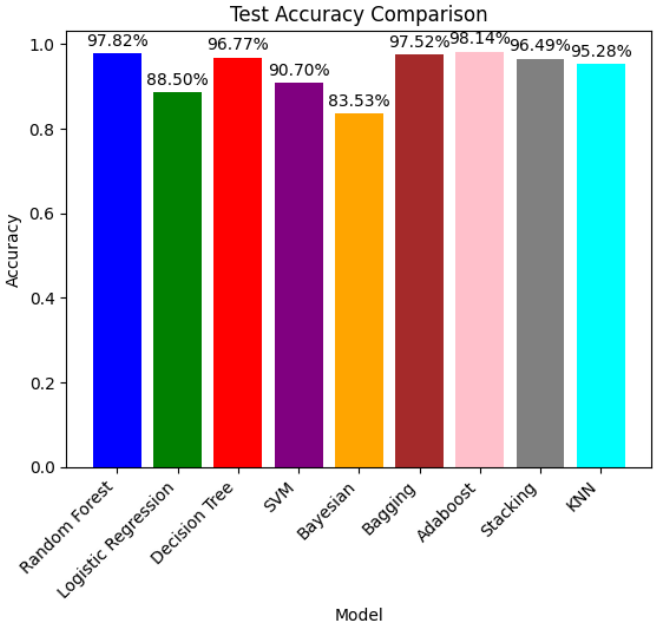

Fig. 19. Comparative testing accuracy across models.

machine learning classification situations where the output might include two or more classes. These metrics were determined using Eq. (3) and Eq. (4). Four distinct combinations of expected and actual values are shown in the Table I.

The Accuracy is calculated using the following formula:

$$\text{Accuracy} = \frac{TP + TN}{T} \tag{3}$$

The F1-score is calculated using the following formula:

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

TABLE I. CONFUSION MATRIX WORKING STRATEGY

		Predicted Class		
		Yes	No	Total
Actual Class	Yes	TP	FN	P
	No	FP	TN	N
	Total	P'	N'	P+N

The terms used in the formulas are as follows:

- TP: True Positive
- FP: False Positive
- TN: True Negative
- FN: False Negative
- T: Total number of samples

C. ROC Curve

In this experiment, we utilized the curve of receiver operating characteristics (ROC) as well as the area under the curve (AUC) parameters for evaluating the effectiveness of around ten machine learning classification algorithms for binary classification tasks. ROC curve analysis was used to evaluate classification, and it is valid at various decision thresholds and sheds light on the trade-off between the percentage of false positives (1-specificity) and the positive rate (sensitivity). To assess each classifier's overall discriminatory power, the area under the curve (also known as the metric was utilized. Consequently, we discover that SVM has powerful discriminative power with an AUC of 0.91, but lower than all other models. RF, bagging, and our suggested model (AB) have perfect discriminative power, obtaining an AUC of 1.00. This suggests that it can successfully discriminate between positive and negative examples in our sample.

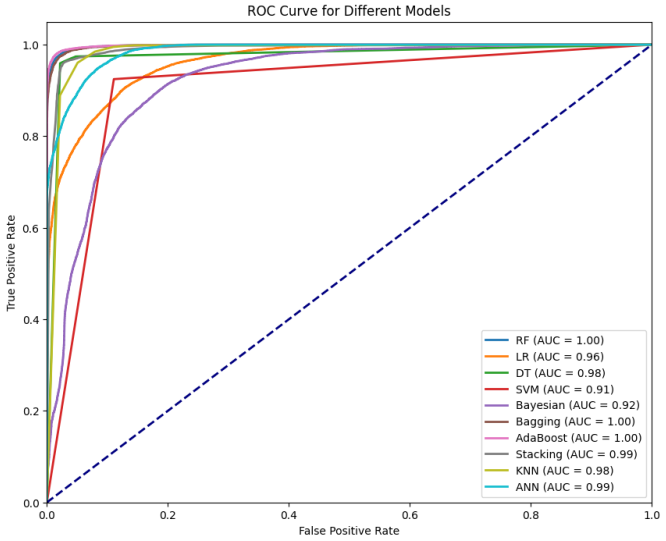

Fig. 20. An ROC curve for showing the performance of all classification model.

D. Model Evaluations

The model's particular classification accuracy is displayed in Table II. These models include several methods, each of which provides a different method for resolving categorization problems. The Proposed Moderated Ada-boost (AB), among them, performs well in testing, with an accuracy of around 98.14%. The Random Forest model retains an astonishing 97.82% accuracy during testing while having a training accuracy of 99.95%. By comparison, the accuracy of the Naive Bayes (NB) model is lower; it recorded an estimated 83.62% in training and a slightly better 83.53% in testing. With a 96.53% training accuracy and a promising 95.28% testing accuracy, the K Nearest Neighbor (KNN) approach performs admirably. Comparably, the Bagging Classifier (BC) model performs admirably, achieving a training accuracy of 99.67% and testing accuracy of 97.52%. The Random Forest model achieves a balanced accuracy of 99.95% in training and a little lower 97.82% in testing, placing it in close alignment with the Proposed Moderated Ada-boost (AB) model. However, even if

the Bagging Classifier performs better in training than other models, it falls well short in testing, an accomplishment that has already been discussed. This vast amount of data helps choose the best model for a given classification assignment by illuminating the strengths and weaknesses of each method. So our Proposed Moderated Ada-boost(AB), provides the highest accuracy among all the applying models.

TABLE II. ACCURACY OF DIFFERENT TYPES OF EVALUATION METRICS

Metrics	Models								
	RF	LR	DT	SVM	NB	BG	ST	AB	KNN
Train Acc	**99.95**	88.50	99.21	90.61	83.62	99.67	98.74	**99.95**	96.53
Test Acc	97.82	88.50	96.77	90.70	83.53	97.52	96.49	**98.14**	95.28
Precision	97.82	88.50	96.78	90.76	84.13	97.52	96.49	**98.15**	95.48
Recall	97.82	88.51	96.78	90.69	83.57	97.51	96.49	**98.14**	95.26
F1 Score	97.82	88.50	96.77	90.70	83.46	97.52	96.49	**98.14**	95.27

Fig. 20 compares the achievement of binary classification approaches that determine whether or not an individual has diabetes in their body using a roc curve. It therefore becomes simple to determine which model is operating at peak efficiency. After comparing ten predictive models at every level, we stumbled upon the following models: Random Forest, Bagging, and Proposed model named Moderated Ada-boost(AB), which perform exceedingly well. Their AUC of 1.00 implies they can successfully distinguish between favorable and adverse occurrences in our dataset. Stacking and Ann come next. According to the basis of our data, nevertheless, the Bayesian classifier achieved an adequate degree of unlawful power with an AUC of 0.92.

V. CONCLUSION AND FUTURE WORK

As a result, a deeper understanding of the opportunities and challenges through enhanced methodology, ethical considerations, and methodological integration will be possible. This research project, which makes use of the "Diabetes_Prediction" dataset, wraps up. Optimization came after the preprocessing-based quality check of the data. Advanced algorithms such as Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Knearest Neighbors (KNN), Naive Bayes (NB), Stacking Classifiers (ST), Bagging Classifier (BG), and Moderated Ada-boost(AB) were utilized in the development of diabetes detection models. The appropriate assessment of measures like accuracy, precision, loss, and F1 score to get the intended performance determines how effective this strategy is. Throughout every phase of this research, there were ethical requirements to maintain confidentiality and handle patient data responsibly. The difficulties in interpreting imbalanced datasets provide new avenues for investigation and creativity. In the end, diabetes diagnosis and machine learning can advance sustainable healthcare, empower patients, and enhance the delivery of medical care.

The application of deep learning and machine learning techniques brings up several possibilities for further research and development in the precise diagnosis of diabetes. Here are a few potential prospects in the future:

- Application to other diseases: Other diseases can be diagnosed using the methods that were developed and accepted for the diagnosis of diabetes. By identifying different human disorders, one may play a special role in the healthcare industry.
- Combining Multiple Data Modes: This model was created taking into account the various physical circumstances that exist among individuals. Further advancements in healthcare might be made feasible by the collecting and integration of diverse physical condition data from several sensors using IOT devices.
- Real-Time Disease Monitoring: Creating technologies that allow patients to simply keep updated about their physical health in real-time. and can thus receive immediate alerts.
- Mobile and Web Applications: Creating user-friendly mobile and web applications that allow patients to create disease reports by entering details about their physical conditions and offering a real-time, graphical user interface that offers management recommendations for diseases.
- Disease prognosis and early warning system: The development of prediction models that can anticipate disease outbreaks based on environmental and historical data is necessary for disease prognosis and early warning systems.
- Patient-doctor communication: If the patient shares information about their physical state, the doctor can use that knowledge to prescribe actions that will help the condition, allowing the patient to take control of their own care.

All things considered, these projects offer a promising direction for further study and development to improve diabetic illness detection techniques and their usefulness.

ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to the almighty Allah who offered our family and us kind care throughout this journey. Also, we wish to express our sincere thanks to our guide, Most. Jannatul Ferdous, Assistant Professor, Department of Computer Science and Engineering for allowing us to work under her on the project. We truly appreciate and value her esteemed guidance and encouragement from the beginning to the end of this project. We are extremely grateful to her. We want to thank to all our teachers for providing a solid background for our studies and research thereafter. They have been a great source of inspiration to us and we thank them from the bottom of my heart. We also want to thank our parents, who taught us the value of hard work by their example. We would like to share this moment of happiness with our parents. They rendered us enormous support during the whole tenure of our stay at Bangladesh University of Business Technology (BUBT). Finally, we are grateful to all our faculty members of the CSE department, BUBT, for making us compatible to complete this research work with the proper guidance and support throughout the last four years.

REFERENCES

[1] K. Ogurtsova, J. da Rocha Fernandes, Y. Huang, U. Linnenkamp, L. Guariguata, N. H. Cho, D. Cavan, J. Shaw, and L. Makaroff, "Idf diabetes atlas: Global estimates for the prevalence of diabetes for 2015 and 2040," *Diabetes research and clinical practice*, vol. 128, pp. 40–50, 2017.

[2] O. M. Disdier-Flores, L. A. Rodríguez-Lugo, R. Pérez-Perdomo, and C. M. Pérez-Cardona, "The public health burden of diabetes: a comprehensive review," *Puerto Rico Health Sciences Journal*, vol. 20, no. 2, 2013.

[3] J. E. Shaw, R. A. Sicree, and P. Z. Zimmet, "Global estimates of the prevalence of diabetes for 2010 and 2030," *Diabetes research and clinical practice*, vol. 87, no. 1, pp. 4–14, 2010.

[4] W. H. Herman, "The global burden of diabetes: an overview," *Diabetes mellitus in developing countries and underserved communities*, pp. 1–5, 2017.

[5] M. J. Uddin, M. M. Ahamad, M. N. Hoque, M. A. A. Walid, S. Aktar, N. Alotaibi, S. A. Alyami, M. A. Kabir, and M. A. Moni, "A comparison of machine learning techniques for the detection of type-2 diabetes mellitus: Experiences from bangladesh," *Information*, vol. 14, no. 7, p. 376, 2023.

[6] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019.

[7] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques," in *2019 1st International informatics and software engineering conference (UBMYK)*. IEEE, 2019, pp. 1–4.

[8] D. Dutta, D. Paul, and P. Ghosh, "Analysing feature importances for diabetes prediction using machine learning," in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2018, pp. 924–928.

[9] P. Sonar and K. JayaMalini, "Diabetes prediction using different machine learning approaches," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2019, pp. 367–371.

[10] M. Soni and S. Varma, "Diabetes prediction using machine learning techniques," *International Journal of Engineering Research & Technology (Ijert) Volume*, vol. 9, 2020.

[11] S. Saru and S. Subashree, "Analysis and prediction of diabetes using machine learning," *International journal of emerging technology and innovative engineering*, vol. 5, no. 4, 2019.

[12] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *Ict Express*, vol. 7, no. 4, pp. 432–439, 2021.

[13] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science*, vol. 132, pp. 1578–1585, 2018.

[14] N. Yuvaraj and K. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster," *Cluster Computing*, vol. 22, no. Suppl 1, pp. 1–9, 2019.

[15] J. Ramesh, R. Aburukba, and A. Sagahyroon, "A remote healthcare monitoring framework for diabetes prediction using machine learning," *Healthcare Technology Letters*, vol. 8, no. 3, pp. 45–57, 2021.

[16] S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, "Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1. IEEE, 2021, pp. 141–146.

[17] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health information science and systems*, vol. 8, pp. 1–14, 2020.

[18] A. Ashiquzzaman, A. K. Tushar, M. R. Islam, D. Shon, K. Im, J.-H. Park, D.-S. Lim, and J. Kim, "Reduction of overfitting in diabetes prediction using deep learning neural network," in *IT Convergence and Security 2017: Volume 1*. Springer, 2018, pp. 35–43.

[19] N. El-Rashidy, N. E. ElSayed, A. El-Ghamry, and F. M. Talaat, "Utilizing fog computing and explainable deep learning techniques for gestational diabetes prediction," *Neural Computing and Applications*, vol. 35, no. 10, pp. 7423–7442, 2023.

[20] A. Dutta, M. K. Hasan, M. Ahmad, M. A. Awal, M. A. Islam, M. Masud, and H. Meshref, "Early prediction of diabetes using an ensemble of machine learning models," *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, p. 12378, 2022.

[21] H. El Massari, Z. Sabouri, S. Mhammedi, and N. Gherabi, "Diabetes prediction using machine learning algorithms and ontology," *Journal of ICT Standardization*, vol. 10, no. 2, pp. 319–337, 2022.

[22] R. Krishnamoorthi, S. Joshi, H. Z. Almarzouki, P. K. Shukla, A. Rizwan, C. Kalpana, B. Tiwari *et al.*, "A novel diabetes healthcare disease prediction framework using machine learning techniques," *Journal of healthcare engineering*, vol. 2022, 2022.

[23] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106773, 2022.

[24] O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. Flores Castañeda, and M. Cabanillas-Carbonell, "Application of machine learning models for early detection and accurate classification of type 2 diabetes," *Diagnostics*, vol. 13, no. 14, p. 2383, 2023.

[25] M. Al-Tawil, B. A. Mahafzah, A. Al Tawil, and I. Aljarah, "Bio-inspired machine learning approach to type 2 diabetes detection," *Symmetry*, vol. 15, no. 3, p. 764, 2023.

[26] R. Shah, J. Petch, W. Nelson, K. Roth, M. D. Noseworthy, M. Ghassemi, and H. C. Gerstein, "Nailfold capillaroscopy and deep learning in diabetes," *Journal of Diabetes*, vol. 15, no. 2, pp. 145–151, 2023.

[27] G. Annuzzi, A. Apicella, P. Arpaia, L. Bozzetto, S. Criscuolo, E. De Benedetto, M. Pesola, R. Prevete, and E. Vallefuoco, "Impact of nutritional factors in blood glucose prediction in type 1 diabetes through machine learning," *IEEE Access*, vol. 11, pp. 17 104–17 115, 2023.

[28] J. J. Sonia, P. Jayachandran, A. Q. Md, S. Mohan, A. K. Sivaraman, and K. F. Tee, "Machine-learning-based diabetes mellitus risk prediction using multi-layer neural network no-prop algorithm," *Diagnostics*, vol. 13, no. 4, p. 723, 2023.

[29] A. R. Kulkarni, A. A. Patel, K. V. Pipal, S. G. Jaiswal, M. T. Jaisinghani, V. Thulkar, L. Gajbhiye, P. Gondane, A. B. Patel, M. Mamtani *et al.*, "Machine-learning algorithm to non-invasively detect diabetes and prediabetes from electrocardiogram," *BMJ Innovations*, vol. 9, no. 1, 2023.

[30] C. J. Ejiyi, Z. Qin, J. Amos, M. B. Ejiyi, A. Nnani, T. U. Ejiyi, V. K. Agbesi, C. Diokpo, and C. Okpara, "A robust predictive diagnosis model for diabetes mellitus using shapley-incorporated machine learning algorithms," *Healthcare Analytics*, vol. 3, p. 100166, 2023.

[31] K. V. Narayan, E. W. Gregg, A. Fagot-Campagna, M. M. Engelgau, and F. Vinicor, "Diabetes—a common, growing, serious, costly, and potentially preventable public health problem," *Diabetes research and clinical practice*, vol. 50, pp. S77–S84, 2000.