Global Journal of Engineering and Technology Advances
eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/

(RESEARCH ARTICLE)

Check for updates

Leveraging machine learning for diabetes prediction: Ensemble model

McDonald Otieno Ogutu [1,*], Benson Nzioka Kituku [2] and Simon M. Karume [3]

[1] *Department of Computer Science and Information Technology, Cooperative University of Kenya, Nairobi, Kenya.*
[2] *School of Computer Science and Information Technology, Dedan Kimathi University, Nyeri, Kenya.*
[3] *School of Science, Engineering and Technology, Kabarak University, Nakuru, Kenya.*

Global Journal of Engineering and Technology Advances, 2025, 25(01), 142-155

Publication history: Received on 02 August 2025; revised on 04 October 2025; accepted on 07 October 2025

Article DOI: https://doi.org/10.30574/gjeta.2025.25.1.0267

Abstract

Diabetes presents great global health challenge, with delayed diagnosis significantly impeding effective management, particularly in resource-constrained regions. This project aimed to enhance timely and accurate diabetes prediction by developing an advanced ensemble machine learning model. A hybrid dataset, compiled from the PIMA Indian (768 instances) and Hospital Frankfurt Germany (2000 instances) datasets, totaling to 2768 datapoints, was utilized to improve generalizability beyond single-source limitations. The methodology involved comprehensive data preprocessing, including the critical imputation of physiologically impossible zero values and feature standardization. F1-score was selected as the primary performance metric due to its ability to provide a vital balance between precision and recall, which is crucial in a medical context where both false positives and false negatives carry significant consequences. Six single classifier models—Logistic Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, and XGBoost—were trained on the data and evaluated after hyperparameter tuning. The F1-scores of these optimized models were: Logistic Regression (0.6328), Decision Tree (0.9843), K-Nearest Neighbors (0.9869), Support Vector Machine (0.9843), Random Forest (0.9947), and XGBoost (0.9974). Based on these results, XGBoost and Random Forest were selected as base learners for a Stacking Classifier ensemble, which utilized a Logistic Regression meta-learner. The developed ensemble model demonstrated exceptional performance, achieving near-perfect ROC-AUC of 0.9999 and an F1-score of 0.9974. This performance not only surpassed results from recent studies but also highlighted the significant potential of machine learning to predict diabetes accurately. The project recommended further development and integration of the ensemble model into a web application.

Keywords: Machine learning; Support vector machine; Gradient boosting; Random Forest; Decision Tree

1. Introduction

Diabetes is a disease affecting many people globally, causing serious health problems ((WHO), 2021). Diabetes must be detected early and accurately to be treated effectively. To respond to this, in the recent decade, data science has come up with powerful machine learning tools in the healthcare sector, providing innovative disease prediction and management solutions. This project explores the possibility of leveraging advanced ensemble machine learning classifiers to improve diabetes prediction, with the goal of increasing accuracy, reducing misdiagnosis, and ultimately contributing to better health outcomes.

Diabetes ranks among the top prevalent diseases globally. World Health Organization ((WHO), 2021), asserts that this condition's prevalence among adults of over 18 years is 8.5% and has caused 6.7 million deaths worldwide in 2021. The disease accounts for a substantial portion of premature deaths, alongside cardiovascular conditions, cancer, and respiratory diseases. Despite a decline in diabetes-related deaths from 2000 to 2010, statistics show a resurgence between 2010 and 2016, with mortality rates expected to increase further ((WHO), 2021). The disease also comes along

* Corresponding author: McDonald Otieno Ogutu

with huge health cost implications. In 2021, the disease caused a global health spending of at least US $960 billion. In Sub-Saharan Africa, the burden of diabetes is expected to impact significantly, with its prevalence anticipated to increase 2.5 times between 2021 and 2045. The spending in health related to diabetes is expected to go up from a total of 12.6 billion USD to 46.7 billion USD ((IDF), 2021). From the year 2015, diabetes has been treated with a lot of concern in Kenya. The disease burden in Kenya has been exacerbated by late diagnosis (Manyara, Mwaniki, Gill, & Gray, 2024). A national survey conducted in 2015 revealed that 51% of diabetes cases in urban areas had not diagnosed. Similarly, a cross-sectional study conducted by (Manyara, Mwaniki, Gill, & Gray, 2024) on some 50 patients in Nairobi revealed that 52% were diabetic but had not been diagnosed. This situation is being worsened by the fact that Kenya has a critical shortage of medical professionals, with only 1 doctor available for every 5,263 people according to Kenya National Bureau of Statistics. This is far below the WHO-recommended ratio of 1:1,000. This shortage severely hampers timely diagnosis and management of chronic diseases such as diabetes. As a result, many cases remain undetected until complications arise, contributing to increased morbidity and preventable mortality. The gap in early screening and diagnosis is especially alarming given the rising burden of diabetes in the country. There is a pressing need for innovative, scalable tools—such as machine learning models—that can assist in early prediction and intervention, especially in underserved regions where access to qualified healthcare providers is limited.

In healthcare, machine learning techniques can be employed to aid in detecting the disease early enough hence its treatment. The technique analyzes the medical dataset to predict results hence lower costs of identifying complex diseases (Gadekallu, et al., 2020). By incorporating machine learning approaches, researchers and studies have succeeded in building machine learning models that are able to detect diabetes early enough, hence avoiding severe effects of the disease and ensuring early medication. Despite this breakthrough, a good percentage of these models are single classifiers which are prone to overfitting in cases of small datasets and this results into poor generalizability. Another limiting factor with single classifiers is that they are affected by noise and outliers, struggle when it comes to bias-variance tradeoff and they are not robust enough to capture complex patterns. In bid to improve predictive capabilities of single classifier machine learning models in diabetes, recent studies have gone the ensemble way. Compared to single machine learning models, ensemble models do better on accuracy, flexibility and high predictive capabilities. However, majority of the studies and researchers who have developed ensemble models classifiers to predict the risk of the disease have mostly used PIMA Indian dataset. This dataset comes with limitations such as the size; it only has 768 instances. Secondly, it represents a given ethnic group thus devoid of diversity hence not able to consider the different genetic predispositions to diabetes present in other population characteristics like the Europeans and Africans. To add on, using PIMA dataset excludes considerations such as environmental factors and lifestyle which largely influences the risk of diabetes. Lastly, because of the nature of this dataset, it limits generalizability and adoption in healthcare globally.

It is on this basis that this project sought to develop from a hybrid of datasets (PIMA (768 datapoints) and Hospital Frankfurt Germany (2000 datapoints)), an accurate ensemble machine learning algorithm from single machine learning models to enhance diabetes prediction. Two best performing models from Random Forest, Support Vector Machines, XGBoost, and k-nearest neighbor were used to develop the ensemble model.

1.1. Study Objectives

1.1.1. General objective

This project's primary goal is to develop an ensemble machine-learning model for predicting diabetes.

1.1.2. Specific objectives

- To review existing machine learning models used in diabetes prediction.
- To evaluate the performance of Logistic regression, XGBoost, Decision trees, SVM, K-NN and Random Forest, in predicting diabetes cases
- To determine the best two machine learning models that are highly accurate in predicting diabetes
- To develop an ensemble ML model from the two best performing ML algorithms
- To evaluate the performance of the resultant ensemble ML model compared to single classifiers

1.2. Significance of the study

This study holds an important place in public health, specifically in diabetes prediction. This research study plays part in developing a diagnostic model for predicting diabetes thus helping in early detection of diabetes which in turn allows for proper medication and management of the disease. This goes a long way in reducing the burden and cost associated with the disease.

A meta-analysis of machine learning models presents the medical field with valuable insights as to which machine learning approaches to adopt after considering the strengths and weaknesses of each. This will give researchers and medical professionals guidance when choosing appropriate machine learning models to use in predicting diabetes.

The predictive accuracy of the developed ensemble machine learning model from best performing single classier models is expected to improve the accuracy significantly. This innovation can lead to more robust and flexible prediction systems that can adapt to diverse datasets and populations, enhancing the generalizability of the models.

As a country, in the spirit of primary healthcare where the government of Kenya is focusing more on preventive approaches to health as opposed to curative approaches, this innovation will come in handy in helping through screening of members of households at the community level. This will lead to early detection and medication thus preventing the disease from progressing to advanced stages leading to out-of-pocket catastrophic costs for the families.

2. Literature review

Lately, following the enormous breakthrough in the machine learning domain, learning techniques have been devised which are quite good at enhancing the detection of cases of diabetes mellitus. In ensemble learning, one has a learning technique, known as the super learner, increasing accuracy by combining outputs from different machine learning algorithms. In their paper, (Dogru, Buyrukoglu, & Ari, 2023) developed the super learner framework based on four classifier models: LR, DT, RF, and gradient boosting; and a meta-learning component using SVM. The study evaluated this model on three datasets to confirm its efficacy. The findings proved that the super learner system performed very well in comparison with each base learner model as a fore-runner of the high-accuracy system for diagnosing diabetes: early-stage risks were forecast with 99.6% accuracy, PIMA data at 92%, and diabetes data from 130 US hospitals at 98%.

In their paper, (Dutta, et al., 2022) put forward a new dataset for diabetes from Bangladesh and an automated classification pipeline with a weighted ensemble of machine learning classifiers, namely NB, RF, DT, XGBoost, and LightGBM. It implements Grid Search hyperparameter optimization on K-fold cross-validation, critical hyperparameters, feature selection, and missing value imputation. The results indicated a statistically significant improvement in diabetes prediction performance with the proposed weighted ensemble (Decision Tree + Random Forest + XGBoost + Light GBM) along with preprocessing techniques to attain 0.735 for accuracy and 0.832 for AUC. The study further showed that statistical imputation and RF-based feature selection combined with the identified ensemble technique had given the best results for predicting the risk of diabetes early enough.

Employing Bayesian networks and radial basis function, (Mahesh, et al., 2022) came up with a blended ensemble machine learning model to aid in diabetes prediction. They used the developed ensemble model to compare the performance of LR, DT, SVM, K-NN and RF. The study found that the resultant developed ensemble model performed better than the traditional single-learning classifiers by achieving an accuracy score of 97.11%.

According to (Atif, Anwer, & Talib, 2022), they opine that single classifier models come with several limitations, the main one being compromising on accuracy because of the models' lack of generalizability over different datasets. In remedying the situation, they developed a model employing hard voting classifier by combining SVM, LR and DT algorithms. In evaluating their model, they used the PIMA dataset and the Early-Stage Diabetes Risk Prediction Dataset. The results indicated that the ensemble model had superior outcomes with accuracies of 81.17% on PIMA dataset and 94.23% on Early-Stage Diabetes Risk Prediction Dataset. The conclusion from this study was that the ensemble models based on hard voting classifiers especially enhanced prediction in terms of accuracy and reliability different sets of datasets notwithstanding (Atif, Anwer, & Talib, 2022).

In another research, (Abnoosian, Farnoosh, & Behzadi, 2023) utilizing imbalanced Iraqi Patient Data, employed a pipeline-based multi-classification approach in prediction of diabetes in three unique groups: the non-diabetic, prediabetes and the diabetic. The study used base classifiers like Gaussian Naive Bayes (GNB), k-NN, RF, SVM, AdaBoost and DT. Since the dataset was imbalanced, they used a weighted ensemble method anchoring on the Area Under the Receiver Operating Characteristic Curve (AUC). To optimize performance, the study employed grid search and Bayesian optimization for hyper-parameter tuning. The resultant ensemble machine learning model compared to single machine learning models, showed superior performance. It gave an accuracy of 0.9887, F1-score of 0.9851, precision of 0.9861, a recall of 0.9792 and an AUC of 0.999.

A study by (H. Qi, 2023) developed an ensemble learning model which they named KFPredict. This model incorporated various input models with significant features and different single classifier models for diabetes prediction. The model

was developed by creating neural network model (KF_NN), that was multi-input in nature. It employed recursive feature in decision trees elimination algorithm together with correlation method to determine significant and non-significant variables. The KF-NN model was then infused with KNN, SVM and RF for the purpose of soft voting and, hence creating a predictive model. The results showed that KFPredict attained accuracy score of 93.5%, sensitivity score of 85% and specificity score of 98%. Compared to single classifier models, this was an improvement of 18.18%. This research underscored the effectiveness of the KFPredict approach in giving robust prediction outcomes on PIMA diabetes data (H. Qi, 2023).

An ensemble machine learning algorithm was modelled by (A. Singh, 2021) to predict diabetes employing XGBoost, RF, SVM, Neural Network, and DT. They named it eDiaPredict. Different evaluation metrics like specificity, Gini Index, minimum error rate, area under the curve (AUC), accuracy, area under the convex hull, sensitivity and minimum weighted coefficient were used to evaluate its performance. PIMA Indian dataset was used and eDiaPredict achieved an accuracy score of 95%. According to this study, eDiaPredict enhanced the prediction of diabetes through ensemble modeling.

Another voting classifier known as En-RfRsK developed by (Amma, 2024) integrated three single ML classifiers namely, K-Nearest Neighbor, Random Forest and Radial SVM to predict the risk of diabetes. To evaluate this ensemble model, the PIMA data was used. The results showed that En-RfRsK was superior to single classifiers by achieving an accuracy score of 88.89% thus showing its usefulness and effectiveness in enhancing the predictive performance.

In predicting blood sugar levels (H. Yang, 2024) used an enhanced stacking ensemble approach that incorporated three enhanced Long Short-Term Memory (LSTM) network models like single classifiers. To ensure adaptive weighting, the study used improved Nearest Neighbor Propagation Clustering Algorithm. To evaluate the model's performance, the study used the OhioT1DM dataset. Results showed that this model achieved a Root Mean Square Error (RMSE) of 1.425 mg/dL, Mean Absolute Error (MAE) of 0.721 mg/dL, and Matthews Correlation Coefficient (MCC) of 0.982 for a 30-minute prediction horizon. The model achieved RMSEs of 3.212 mg/dL and 6.346 mg/dL, MAEs of 1.605 mg/dL and 3.232 mg/dL, and MCCs of 0.950 and 0.930 for 45-minute and 60-minute horizons respectively. The study concluded that LSTM ensemble technique improved RMSE & MAE by 27.92% and 65.32%, in that order compared to non-ensemble model (H. Yang, 2024).

2.1. Gaps identified in previous works

A review of recent studies on machine learning models for diabetes prediction reveals that a wide range of algorithms—such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, K-Nearest Neighbors, Gradient Boosting, and Naive Bayes have been widely applied. The PIMA Indian Diabetes dataset remains the most frequently used data source across the literature, appearing in studies by (Dogru, Buyrukoglu, & Ari, 2023), (Mahesh, et al., 2022), (Atif, Anwer, & Talib, 2022), (H. Qi, 2023), (A. Singh, 2021), and (Amma, 2024). Despite the variety of algorithms explored, a dominant limitation is the over-reliance on relatively small datasets, particularly the PIMA dataset which only contains 768 instances. This limits the generalizability and robustness of model performance in real-world, diverse populations.

In addition to small sample sizes, other studies such as (Dutta, et al., 2022) and (Abnoosian, Farnoosh, & Behzadi, 2023) have used region-specific datasets like the DDC Bangladesh dataset and the Iraqi Patient Dataset. However, these also come with limitations. For instance, the Iraqi dataset is significantly imbalanced, with a disproportionately higher number of non-diabetic cases (91,500) compared to diabetic ones (8,500), posing challenges for fair model training and evaluation.

2.2. Conclusion from the review

From the literature review, studies on diabetes prediction predominantly rely on relatively small and often imbalanced datasets most notably the PIMA Indian dataset with only 768 instances. This limitation constrains the development of robust and generalizable machine learning models. Additionally, while a variety of algorithms such as Random Forest, Support Vector Machines, XGBoost, and Decision Trees have been tested, few studies have explored ensemble approaches built from multiple strong-performing base learners, especially using diverse data sources.

It is on this basis that this project seeks to bridge these gaps by developing an accurate ensemble machine learning model for diabetes prediction. The model will be built from a hybrid of datasets the PIMA dataset and the Hospital Frankfurt Germany dataset (with over 2,000 datapoints) to improve data diversity and model generalizability. From among the single machine learning models (Random Forest, Support Vector Machines, XGBoost, and K-Nearest Neighbor), the two best-performing algorithms will be selected and combined to form the ensemble model. This

approach aims to enhance predictive accuracy and support early diagnosis, especially in resource-limited settings where delayed detection contributes to rising diabetes-related morbidity and mortality.

3. Methodology

This chapter covers dataset description, the data pre-processing steps, model development, model evaluation metrics, and the selection process for the best models.

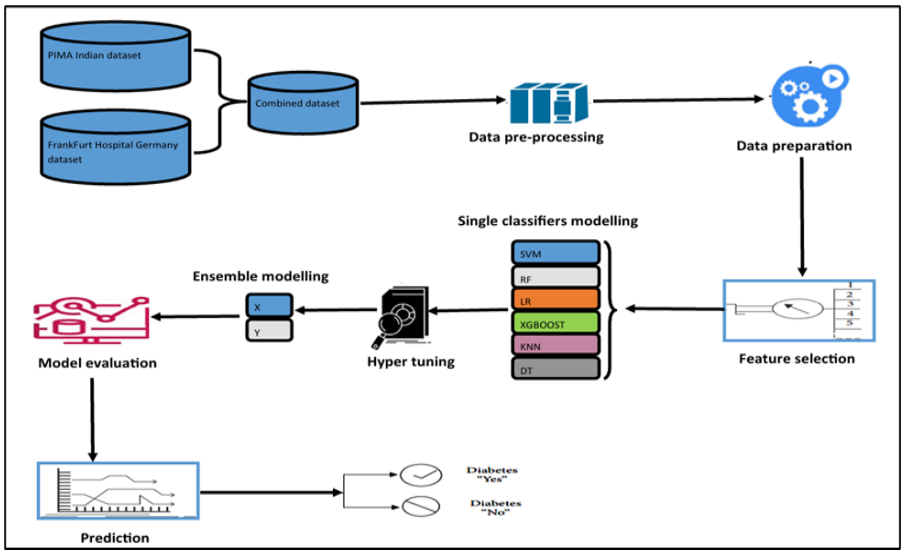

Figure 1 Approach flow for the ensemble modelling

3.1. Data source and description

Aggregated data was sourced from two sources; PIMA Indian dataset (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database) and Dataset of diabetes, taken from the hospital Frankfurt, Germany (https://www.kaggle.com/datasets/johndasilva/diabetes) all of which are available in Kaggle. The Indian PIMA dataset has 768 data points while the Hospital Frankfurt dataset has 2000 data points. The variables in the datasets are diabetes status, glucose level, blood pressure, skin thickness, insulin level, BMI (body mass index), pregnancy, diabetes pedigree function and age. Table 1 below shows the attributes.

Table 1 Data attributes

No	Attribute	Data type	Description
1	Pregnancy	Numeric	The number of pregnancies
2	Glucose	Numeric	Glucose plasma levels two hours after consuming glucose
3	Blood pressure	Numeric	Diastolic blood pressure (mm Hg)
4	Skin thickness	Numeric	Thickness of the skin fold on the triceps of the upper arm (mm)
5	Insulin	Numeric	Insulin serum levels in the blood two hours after the glucose test (lh/ml)
6	BMI	Numeric	Body mass index [weight in kg/(Height in m)]
7	Diabetes Pedigree Function	Numeric	Numerical value that estimates a person's genetic risk of developing diabetes based on their family history.
8	Age	Numeric	Patient age
9	Diabetes status	Categorical	Diagnostic results (Diabetic or not diabetic)

3.2. Data pre-processing steps

3.2.1. Dealing with missing values

Measures of central tendency (median) technique of imputation was used to manage missing values. This approach ensures that the dataset remains comprehensive and usable.

3.2.2. Data standardization

The project employed Min-Max scaling method to ensure that the range of features are normalized. This will ensure that data falls in a given range, always between 0 and 1. This in turn strengthened the robustness of machine learning models that are sensitive to feature scaling.

3.2.3. Encoding

One-hot encoding was employed to assign numerical values to each category of a qualitative variable. This applied to the dependent variable only as it was the only categorical variable in the dataset. The response variable was diabetes status and was dichotomous which is “diabetic” or “not diabetic”. “Diabetic” status was coded as 1 and “not diabetic” status was coded as 0.

3.2.4. Feature selection

Feature selection was conducted to identify the most relevant independent variables for model training and to address potential issues of multicollinearity. Correlation heatmaps were employed to assess the relationships between independent variables. During this assessment, no significant cases of multicollinearity were identified among the independent variables as can be observed in figure 2, therefore, all independent variables were retained for subsequent analysis, ensuring the preservation of their individual predictive utility.

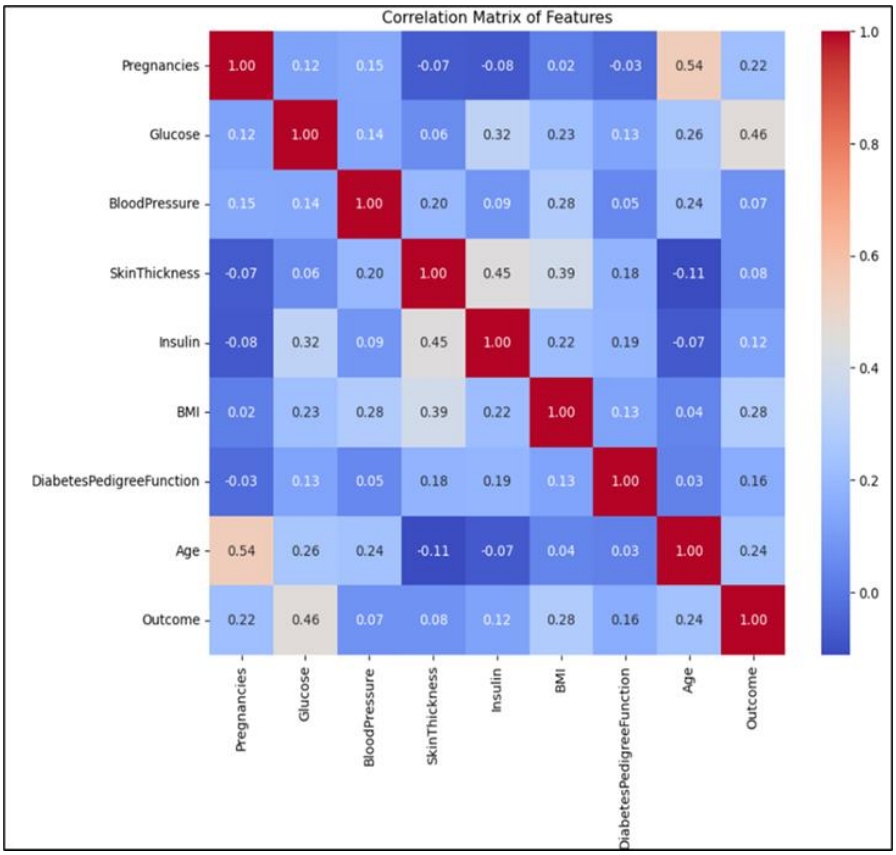

Figure 2 Correlation matrix of features

3.2.5. Class balance analysis

To gain a comprehensive understanding of the dataset's core, the distribution of the target variable (diabetes status) was analyzed, providing a clear picture of the class balance. This step was vital for identifying potential class imbalance that might necessitate specific handling during model training. The diabetics were 65.6% while the non-diabetics were 34.4%. The analysis result is as shown in figure 3 below;

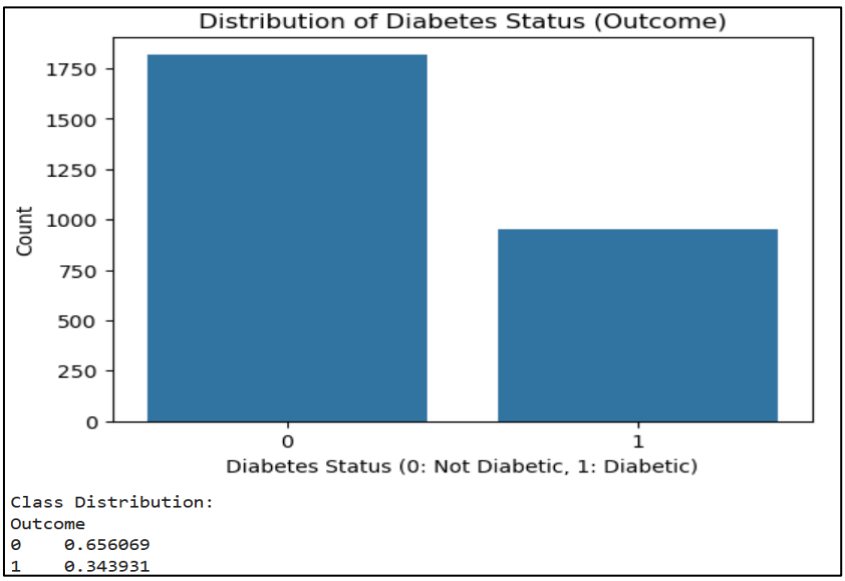

Figure 3 Distribution of diabetes status

3.2.6. Data distribution

The individual characteristics of all variables were explored through histograms, which revealed the distribution patterns of each numerical variable. It can be observed that glucose, blood pressure and BMI were moderately normally distributed. The rest of the variables were skewed. The results were as shown in figure 4 below.

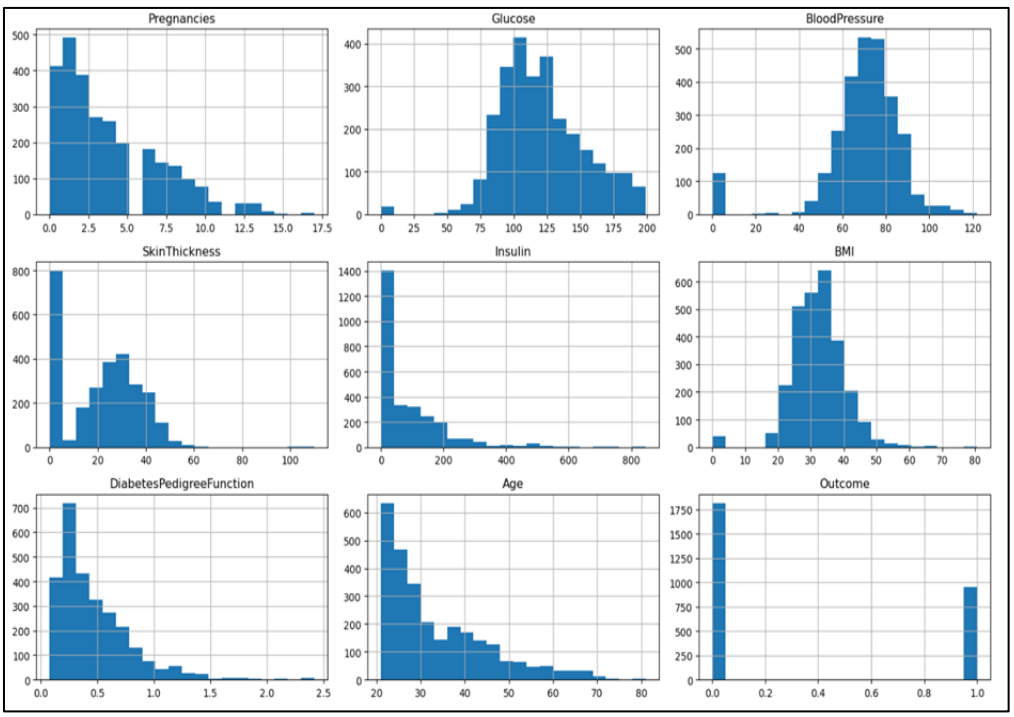

Figure 4 Distribution of variables

3.3. Model building and performance analysis

3.3.1. Model building

The research dataset was divided into two splits: one was the training set, and the other was the testing set. The training set had 80% of the data while the testing set had 20% of the data. The splitting method allows the model to be trained on a majority of this data while retaining a separate set for evaluating its performance, ensuring that the model performs well at generalization. Six single classifier models—Logistic Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, and XGBoost were trained and tested on the data and evaluated after hyperparameter tuning

3.3.2. Performance analysis

To evaluate the performance of each model, the project used several evaluation metrics:

Accuracy: This was to quantify how much of the total number of samples were correctly predicted. It measures the model reliability on classification of a positive or negative cases. On the other hand, when training datasets are imbalanced, accuracy can be misleading; it quickly goes upwards without paying attention to detection of positive class.

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

The subject should be classified as either diabetic or not diabetic. Out of the classification, we had subjects truly classified and falsely classified hence the below;

- **True positive (TP):** The subject is predicted as positive (diabetic) and is diabetic.
- **True negative (TP):** The subject is predicted as negative (not diabetic) and is actually not diabetic.
- **False positive (FP):** The subject is predicted as positive (diabetic) and is not diabetic.
- **False negative (FN):** The subject is predicted as negative (not diabetic) but is actually diabetic.
- **Precision:** Indicates the proportion of true positive rightly predicted among all predicted positives.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}$$

Recall: Reflects the ability of the model to identify all relevant instances (true positives).

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

F1-Score: A harmonic mean of precision and recall, providing a single metric for model evaluation.

Area Under the Receiver Operating Characteristic (ROC) Curve (AUC): Evaluates the model's power to differentiate between positive and negative classes.

3.3.3. Ensemble Machine Learning Development

To leverage the complementary strengths of multiple high-performing individual classifiers and enhance predictive accuracy, an ensemble model was constructed using the Stacking Classifier methodology. This approach involved building a two-layer structure: the first layer consisted of base estimators, which were the two best-performing single classifiers identified previously, namely XGBoost and Random Forest, chosen for their superior F1-scores. The second layer comprised a meta-learner, a Logistic Regression model, which was tasked with combining the predictions generated by these base estimators. During the ensemble's training, a crucial step involved employing 5-fold cross-validation internally, ensuring that the base models generated their predictions on out-of-fold data to prevent data leakage. Furthermore, these base models passed their predicted probabilities to the meta-learner, providing richer information than just hard class labels. The ensemble was configured to utilize all available CPU cores for efficient parallel processing and was set to not pass the original features directly to the final estimator, ensuring the meta-learner solely focused on integrating the insights from the base models' outputs. The complete ensemble model was then trained on the designated training dataset.

4. Results

4.1. Overall performance of single classifier models before and after hyperparameter tuning

The six single classifier models were initially evaluated using their default hyperparameters to establish baseline performance. Subsequently, hyperparameter tuning was performed using GridSearchCV to optimize its predictive capabilities on the Diabetes dataset. It can be observed that there were major improvements in K-Nearest Neighbour and Support Vector Machine in terms of the overall performance after hyperparameter tuning. For example, there was a 48% improvement in recall metric in support vector machine after hyperparameter tuning. The results are as in table 2 below.

Table 2 Model performance comparison before and after hyperparameter tuning

Model	Metric	Default	Tuned	Improvement %
Logistic Regression	Accuracy	0.7726	0.7780	0.70
Logistic Regression	Precision	0.7211	0.7361	2.08
Logistic Regression	Recall	0.5550	0.5550	0.00
Logistic Regression	F1-Score	0.6272	0.6328	0.90
Logistic Regression	ROC-AUC	0.8355	0.8355	0.00
Decision Tree	Accuracy	0.9892	0.9892	0.00
Decision Tree	Precision	0.9843	0.9843	0.00
Decision Tree	Recall	0.9843	0.9843	0.00
Decision Tree	F1-Score	0.9843	0.9843	0.00
Decision Tree	ROC-AUC	0.9880	0.9880	0.00
K-Nearest Neighbor	Accuracy	0.8448	0.9810	17.31
K-Nearest Neighbor	Precision	0.7807	0.9844	26.08
K-Nearest Neighbor	Recall	0.7644	0.9895	29.45
K-Nearest Neighbor	F1-Score	0.7725	0.9869	27.76
K-Nearest Neighbor	ROC-AUC	0.9391	0.9999	6.47
Support Vector Machine	Accuracy	0.8412	0.9892	17.60
Support Vector Machine	Precision	0.8411	0.9843	17.03
Support Vector Machine	Recall	0.6649	0.9843	48.03
Support Vector Machine	F1-Score	0.7427	0.9843	32.53
Support Vector Machine	ROC-AUC	0.9002	0.9996	11.04
Random Forest	Accuracy	0.9964	0.9964	0.00
Random Forest	Precision	1.0000	1.0000	0.00
Random Forest	Recall	0.9895	0.9895	0.00
Random Forest	F1-Score	0.9947	0.9947	0.00
Random Forest	ROC-AUC	0.9999	0.9999	0.00
XGBoost	Accuracy	0.9928	0.9982	0.55
XGBoost	Precision	0.9845	1.0000	1.58
XGBoost	Recall	0.9948	0.9948	0.00

XGBoost	F1-Score	0.9896	0.9974	0.79
XGBoost	ROC-AUC	0.9981	0.9989	0.08

4.2. Performance based on F1-score

In evaluating classification models for diabetes prediction, common metrics like Accuracy, Precision, Recall, and F1-score each offer distinct insights but have limitations. Accuracy can be misleading in imbalanced datasets, as a model might appear effective by simply predicting the majority class. Precision, while indicating the reliability of positive predictions, can lead to numerous missed actual cases (low Recall) if prioritized exclusively. Conversely, Recall, crucial for minimizing dangerous false negatives (missed diagnoses in diabetes), can result in an unacceptably high rate of false positives (unnecessary alarms for healthy individuals) if optimized in isolation. Therefore, to strike a vital balance between accurately identifying true diabetic cases and ensuring the credibility of positive predictions, the F1-score was selected as the primary metric for determining the two best-performing single classifier models. The F1-score effectively captures the harmonic mean of Precision and Recall, making it particularly suitable for this medical diagnostic context where both types of errors carry significant consequences. Table 3 below shows the performance of the 6 single classifier models in terms of F1-score.

Table 3 Models performance comparison by F1-score

Model	F1-Score (Default Parameters)	F1-Score (Tuned Parameters)	Improvement (%)
Logistic Regression	0.6272	0.6328	+0.89%
Decision Tree	0.9843	0.9843	0.00%
K-Nearest Neighbor	0.7725	0.9869	+27.7%
Support Vector Machine	0.7427	0.9843	+32.53%
Random Forest	0.9947	0.9947	0.00%
XGBoost	0.9896	0.9974	+0.79%

As observed in table 3 above, based on the F1-score of the tuned models, the classifiers demonstrated high levels of balanced performance. Among them, XGBoost achieved the highest F1-score of 0.9974, followed closely by Random Forest with an F1-score of 0.9947. These two models, XGBoost and Random Forest, were thus selected as the best-performing single classifiers for subsequent ensemble development.

4.3. Ensemble Model results

The ensemble model achieved an Accuracy of 0.9982, signifying that nearly all predictions were correct. Its Precision was a perfect 1.0000, indicating that every instance predicted as positive for diabetes was indeed a true positive, with no false alarms. The Recall stood at a remarkable 0.9948, meaning the model successfully identified over 99.4% of all actual diabetes cases. These combined strengths resulted in an outstanding F1-Score of 0.9974, reflecting an almost perfect balance between precision and recall. Furthermore, the ROC-AUC of 0.9999 solidified its discriminative power, indicating a near-flawless ability to distinguish between diabetic and non-diabetic individuals across all possible classification thresholds. The confusion matrix for the ensemble model provides a detailed insight into its classification performance:

- **True Negatives (TN):** 363 non-diabetic individuals were correctly identified as non-diabetic.
- **False Positives (FP):** Notably, there were 0 false positive predictions, meaning no healthy individuals were incorrectly diagnosed with diabetes. This is a crucial outcome in a medical context.
- **False Negatives (FN):** Only 1 false negative occurred, indicating that out of all actual diabetic cases, only one was missed by the model.
- **True Positives (TP):** 190 diabetic individuals were correctly identified as diabetic.

The classification report further details the class-wise performance:

- **Class 0 (Non-Diabetic):** The model achieved perfect precision, recall, and F1-score (all 1.00), demonstrating flawless identification of non-diabetic cases.
- **Class 1 (Diabetic):** For the diabetic class, the model also showcased exceptional performance with perfect precision (1.00), nearly perfect recall (0.99), and an F1-score approaching perfection (1.00).

Table 4 Ensemble Model performance

Metric	Performance
Accuracy	0.9982
Precision	1.0000
Recall	0.9948
F1-Score	0.9974
ROC-AUC	0.9999

Table 5 Confusion matrix

	Predicted diabetic	Predicted non-diabetic
Actual diabetic	363	0
Actual non-diabetic	1	190

4.4. Performance comparison between Ensemble model vs single classifier models

The final comparison of all optimized single classifiers and the ensemble model, based on their F1-scores, revealed compelling insights into their balanced predictive capabilities. The Ensemble Model achieved an F1-score of 0.9974, demonstrating an exceptionally high balance between precision and recall. This performance was on par with the XGBoost model, which also achieved an F1-score of 0.9974, positioning both as the top performers.

Following these, the Random Forest model secured a strong F1-score of 0.9947. The K-Nearest Neighbors, Decision Tree, and Support Vector Machine models also exhibited high F1-scores of 0.9869, 0.9843, and 0.9843 respectively. In contrast, the Logistic Regression model, while providing a reasonable baseline, recorded a significantly lower F1-score of 0.6328 compared to the other advanced models.

This comparison underscores the highly effective performance of the ensemble model, confirming its ability to achieve top-tier balanced predictions, matching the excellence of the leading individual classifier, XGBoost. The results are as observed in table 6 below;

Table 6 Ensemble vs single classifier models performance

FINAL PERFORMANCE COMPARISON (Tuned Single Models vs Ensemble Model)					
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
XGBoost	0.9982	1.0000	0.9948	0.9974	0.9989
Ensemble Model	0.9982	1.0000	0.9948	0.9974	0.9999
Random Forest	0.9964	1.0000	0.9895	0.9947	0.9999
K-Nearest Neighbor	0.9910	0.9844	0.9895	0.9869	0.9999
Decision Tree	0.9892	0.9843	0.9843	0.9843	0.9880
Support Vector Machine	0.9892	0.9843	0.9843	0.9843	0.9996
Logistic Regression	0.7780	0.7361	0.5550	0.6328	0.8355

4.5. Results comparative analysis summary

Both the XGBoost and the Ensemble Model demonstrated exceptional and identical F1-scores of 0.9974, indicating a perfect balance between precision and recall at the chosen classification threshold. However, when considering the overall discriminative ability across all possible thresholds, the Ensemble Model marginally surpassed XGBoost, achieving a ROC-AUC of 0.9999 compared to XGBoost's 0.9989. This superior ROC-AUC suggests the Ensemble Model offers a slightly more robust and comprehensive capability in distinguishing between diabetic and non-diabetic cases.

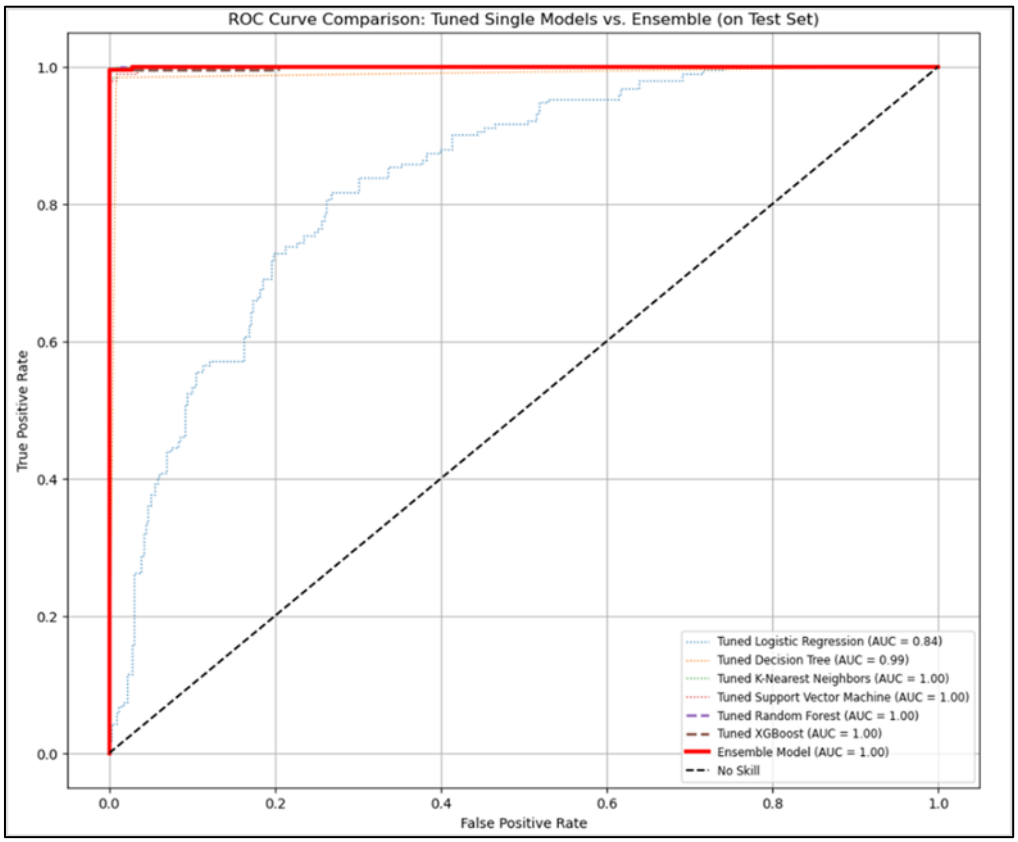

Figure 5 ROC curve comparison: Ensemble vs Single classifier models

5. Discussion

The performance achieved by this study's models, particularly the optimized single classifiers and the ensemble model, compares favorably with and in several instances, surpasses results reported in existing literature on diabetes prediction.

Previous studies using the PIMA Indian dataset (which typically has 768 instances) reported accuracies ranging from 88.89% to 97.11%. For instance, A. Doğru et al. (2023) achieved 92% accuracy, T. Mahesh et al. (2022) reported 97.11%, M. Atif et al. (2022) found 94.23%, H. Qi et al. (2023) 93.5%, A. Singh et al. (2021) 95%, and N. Bhuvaneswari (2024) 88.89%. Our optimized XGBoost model achieved an accuracy of 0.9982 and Random Forest reached 0.9964, both significantly higher than most cited accuracies.

A key aspect of this study was the creation of a hybrid dataset by combining the PIMA Indian dataset (762 instances) with the Hospital Frankfurt Germany dataset (2000 instances), resulting in a larger dataset of 2762 datapoints. This methodology directly addressed the common limitation noted in the literature regarding the relatively small size of datasets (often 768 instances for PIMA). This increased dataset size likely contributed to the enhanced generalizability and superior performance observed in our models. Similar to K. Abnoosian et al. (2023), who reported an accuracy of 98.87% and F1-score of 98.51% on an Iraqi Patient Dataset, this study emphasized F1-score and ROC-AUC, recognizing their value for potentially imbalanced datasets in medical contexts. Our ensemble model achieved an F1-score of 0.9974 and an ROC-AUC of 0.9999, which represents a superior balanced performance and discriminative capability compared to many prior works. The robust performance achieved suggests that comprehensive data preparation (including

handling 'zeros' as missing values) and systematic optimization strategies can yield highly accurate and reliable models even when dealing with combined data sources.

6. Conclusion

Diabetes stands as a global health crisis, responsible for millions of premature deaths and substantial healthcare expenditures. A critical challenge in combating this disease, particularly evident in regions like Kenya, is the high prevalence of delayed diagnoses, significantly exacerbating the disease burden and patient outcomes. Recognizing the limitations of traditional diagnostic methods and the inherent weaknesses of single-classifier machine learning models such as susceptibility to overfitting on small datasets, sensitivity to noise, and limited capacity to capture complex patterns, this project embarked on leveraging advanced ensemble machine learning techniques for early and accurate diabetes prediction.

The primary objective of this project was successfully achieved: the development of a highly accurate ensemble machine learning model for diabetes prediction. Through a structured methodology, the study comprehensively evaluated six widely used individual machine learning classifiers: Logistic Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, and XGBoost. The importance of meticulous data preprocessing was underscored, particularly in handling the 'zero' values in physiological measurements that denote missing data, and through the application of feature scaling. Hyperparameter tuning proved instrumental in optimizing the performance of these base classifiers, significantly enhancing models like K-Nearest Neighbors (F1-score improved from 0.7725 to 0.9869) and Support Vector Machine (F1-score improved from 0.7427 to 0.9843).

Among the individually optimized classifiers, XGBoost (F1-score: 0.9974) and Random Forest (F1-score: 0.9947) demonstrated the highest balanced predictive capabilities. These two robust models were subsequently chosen as the base estimators for the ensemble model. The developed ensemble, a Stacking Classifier utilizing a Logistic Regression final estimator, achieved an exceptional performance on the held-out test set, mirroring the F1-score of 0.9974 observed in XGBoost. Critically, the ensemble model exhibited a remarkable ROC-AUC of 0.9999, marginally surpassing XGBoost's 0.9989. This near-perfect AUC, coupled with its flawless precision (1.0000) and only a single false negative, underscores the ensemble's superior overall discriminative power and its ability to minimize both false alarms and missed diagnoses, a crucial aspect in clinical application.

A pivotal contribution of this study was the creation of a hybrid dataset, combining the PIMA Indian dataset (768 instances) with the Hospital Frankfurt Germany dataset (2000 instances), resulting in a larger and more diverse dataset of 2768 datapoints. This directly addressed a prevalent limitation identified in literature, where many studies rely solely on the smaller, ethnically specific PIMA dataset, thus enhancing the generalizability of the developed models beyond a single ethnic group or limited sample size. The high performance achieved despite the inherent characteristics of the combined public datasets validates the effectiveness of robust preprocessing and ensembles learning strategies.

In conclusion, this project successfully developed an accurate and robust ensemble machine learning model for diabetes prediction. The final ensemble model's exceptional performance, particularly its high F1-score and near-perfect ROC-AUC, demonstrates the significant potential of advanced machine learning techniques to aid in early and accurate diabetes diagnosis, thereby contributing to improved health outcomes and more efficient healthcare resource allocation, especially in regions like Kenya facing critical shortages of medical professionals.

Compliance with ethical standards

Acknowledgments

I would like to express my sincere gratitude to everyone who contributed to the successful completion of this project.

First and foremost, I extend my heartfelt appreciation to my supervisor, Prof. Karume and Dr. Kituku, for their invaluable guidance, continuous support, and constructive feedback throughout the course of this study.

I am also deeply thankful to the faculty and staff of Computer Science & Information Technology Department, whose expertise and resources provided a solid foundation for my research.

Special thanks to my colleagues and friends for their encouragement, insightful discussions, and moral support during challenging moments.

Lastly, I am profoundly grateful to my family for their patience, understanding, and unwavering belief in me throughout this academic journey.

Disclosure of conflict of interest

The authors declare no conflict of interest.

Data Availability Statement

Aggregated data was sourced from two sources; PIMA Indian dataset (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database) and Dataset of diabetes, taken from the hospital Frankfurt, Germany (https://www.kaggle.com/datasets/johndasilva/diabetes) all of which are available in Kaggle.

References

[1] (IDF), I. D. (2021, -- --). *IDF Diabetes Atlas, 10th edition*. (International Diabetes Federation) Retrieved January 23, 2022, from https://diabetesatlas.org/atlas/tenth-edition/

[2] (WHO), W. H. (2021, -- --). *Diabetes*. (World Health Organization) Retrieved January 23, 2022, from World Health Organization: https://www.who.int/health-topics/diabetes

[3] A. Singh, A. D. (2021). eDiaPredict: An ensemble-based framework for diabetes prediction. *ACM Transactions on Multimedia Computing, Communications, and Applications, 17*(2s). doi:10.1145/3415155

[4] Abnoosian, K., Farnoosh, R., & Behzadi, M. H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC Bioinformatics, 24*, 337. doi:https://doi.org/10.1186/s12859-023-05465-z

[5] Amma, N. G. (2024). En-RfRsK: An ensemble machine learning technique for prognostication of diabetes mellitus. *Egyptian Informatics Journal, 25*. doi:10.1016/j.eij.2024.100441

[6] Atif, M., Anwer, F., & Talib, F. (2022). An ensemble learning approach for effective prediction of diabetes mellitus using hard voting classifier. *Indian Journal of Science and Technology, 15*(39), 1978–1986. doi:https://doi.org/10.17485/IJST/v15i39.1520

[7] Dogru, A., Buyrukoglu, S., & Ari, M. (2023). A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. *Medical & Biological Engineering & Computing, 61*(6), 785–797. doi:https://doi.org/10.1007/s11517-022-02749-z

[8] Dutta, A., Hasan, M. K., Ahmad, M., Awal, M. A., Islam, M. A., Masud, M., & Meshref, H. (2022). Early prediction of diabetes using an ensemble of machine learning models. *International Journal of Environmental Research and Public Health, 19*(19), 12378. doi:https://doi.org/10.3390/ijerph191912378

[9] Gadekallu, T. R., Khare, N., Bhattacharya, S., Singh, S., Reddy Maddikunta, P. K., & Ra, I. H. (2020). Early detection of diabetic retinopathy using PCA-Firefly based deep learning model. *Electronics, 9*(3), 274. doi:https://doi.org/10.3390/electronics9030274

[10] H. Qi, X. S. (2023). KFPredict: An ensemble learning prediction framework for diabetes based on fusion of key features. *Computer Methods and Programs in Biomedicine, 231*, 107378. doi:10.1016/j.cmpb.2023.107378

[11] H. Yang, Z. C. (2024). AWD-stacking: An enhanced ensemble learning model for predicting glucose levels. *PLoS ONE, 19*(2). doi:10.1371/journal.pone.0291594

[12] Mahesh, T. R., Kumar, D., Vinoth Kumar, V., Asghar, J., Bazezew, B. M., Natarajan, R., & Vivek, V. (2022). Blended ensemble learning prediction model for strengthening diagnosis and treatment of chronic diabetes disease. *Advances in Public Health*. doi:https://doi.org/10.1155/2022/4451792

[13] Manyara, A. M., Mwaniki, E., Gill, J. M., & Gray, C. M. (2024). Perceptions of diabetes risk and prevention in Nairobi, Kenya: A qualitative and theory of change development study. *PLoS ONE, 19*(2), e0297779. doi:https://doi.org/10.1371/journal.pone.0297779