Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 7, No. 1, February, 2023

A Comparative Performance Analysis of Hybrid and Classical Machine Learning Method in Predicting Diabetes

Kalaiarasi Sonai Muthu Anbananthen [1*], Mikail Bin Muhammad Azman Busst [1], Rajkumar Kannan [2], Subarmaniam Kannan [1]

[1] *Faculty of Information Science and Technology, Multimedia University, Melaka 75450, Malaysia.*

[2] *Bishop Heber College (Autonomous), Tiruchirappalli 620017, India.*

Abstract

Diabetes mellitus is one of medical science's most important research topics because of the disease's severe consequences. High blood glucose levels characterize it. Early detection of diabetes is made possible by machine learning techniques with their intelligent capabilities to accurately predict diabetes and prevent its complications. Therefore, this study aims to find a machine learning approach that can more accurately predict diabetes. This study compares the performance of various classical machine learning models with the hybrid machine learning approach. The hybrid model includes the homogenous model, which comprises Random Forest, AdaBoost, XGBoost, Extra Trees, Gradient Booster, and the heterogeneous model that uses stacking ensemble methods. The stacking ensemble or stacked generalization approach is a meta-classifier in which multiple learners collaborate for prediction. The performance of the homogeneous hybrid models, Stacked Generalization and the classic machine learning methods such as Naive Bayes and Multilayer Perceptron, k-Nearest Neighbour, and support vector machine are compared. The experimental analysis using Pima Indians and the early-stage diabetes dataset demonstrates that the hybrid models achieve higher accuracy in diagnosing diabetes than the classical models. In the comparison of all the hybrid models, the heterogeneous model using the Stacked Generalization approach outperformed other models by achieving 83.9% and 98.5%.

Keywords:

Ensemble Learning;

Stacked Generalization;

Machine Learning;

Prediction;

Healthcare.

Article History:

Received:	12	June	2022
Revised:	20	August	2022
Accepted:	08	September	2022
Available online:	12	October	2022

1- Introduction

Diabetes mellitus is a critical and expensive public health condition that affects people worldwide today. According to the World Health Organization, 171 million people worldwide have diabetes, and that number is expected to rise to 366 million by 2030 [1]. On the other hand, the IDF Diabetes Atlas in 2021 [2] predicted that 537 million adults (20-79 years) have diabetes, which is estimated to reach 783 million by 2045. Saeedi et al. [3] describe the severity of diabetes, estimating that half a billion people globally have diabetes, with rates expected to rise to 25% in 2030 and 51% in 2045. Despite the numbers and estimations varying, diabetes is a serious worldwide issue that greatly impacts individuals worldwide. Although there is no long-term cure for diabetes, it can be treated and prevented with early detection and intervention.

Diabetes occurs when the pancreas cannot produce enough insulin, resulting in an abnormally high blood sugar level. Without insulin, glucose cannot enter body cells and is expelled via urine. Diabetes-related long-term hyperglycemia damages and dysfunctions multiple tissues, including the nerves, kidneys, heart, eyes, and blood vessels. In addition,

[*] **CONTACT**: kalaiarasisma@gmail.com

DOI: http://dx.doi.org/10.28991/ESJ-2023-07-01-08

patients with diabetes have a higher risk of cardiovascular disease [4], lower limb amputation, and even shorter life expectancy than those without diabetes. Also, these issues could result in death. According to the information above, people with diabetes must take particular precautions to avoid complications. Maintaining normal blood glucose levels is a vital precaution. Early detection and management of diabetes can aid in reducing the risk, and one strategy for diagnosing diabetes is through machine learning. Many researchers have used machine learning approaches to diagnose diabetes problems, such as k-Nearest Neighbour, Support Vector Machine, Naive Bayes, Decision Tree, ID3, J48, Decision Table, etc.

This paper compares the performance of different classical and hybrid approaches to find a machine learning approach that can more accurately predict diabetes. The classical models included in this research are Multilayer Perceptron, K-Nearest Neighbor, Support Vector Machine, and Naïve Bayes. Hybrid models are strategies for combining multiple models for prediction. The hybrid model is divisible into homogeneous and heterogeneous models. A homogeneous hybrid model is a learning technique where all the models or classifiers belong to the same type. This research uses Random Forest, Ada Boost, XGBoost, Extra Trees, and Gradient Boosted Tree. The heterogeneous model is a learning technique where all the classifiers belong to different learning types. In this research, we have used a heterogeneous model that uses stacking ensemble methods. The stacking ensemble or stacked generalization approach is where the hybrid learner is created by layering many different base-level classifiers with a meta-classifier. The experimental performance of all methods is compared using a variety of metrics to determine the best method for predicting diabetes.

The primary objectives of this research are:

- To perform a comprehensive performance analysis of classical and hybrid machines (homogeneous models and heterogenous Stacked Generalization);
- To provide an approach for developing stacked Generalization for diabetes prediction;
- To determine the best base and meta-classifier for stacking.

The rest of the paper includes a literature review, proposed approach, results, discussion, and conclusion section.

2- Literature Review

Six people die every minute from diabetes or its complications [5], making it one of the world's most severe health [6] and economic crises. Numerous factors contribute to the disease's progression, and if addressed, personalized therapy profiles can help prevent it from progressing, resulting in lower patient morbidity and mortality rates. As a result, a significant amount of money has been invested in developing intelligent models in this field, focusing on applying machine learning and data mining to diagnose, forecast, and manage the disease [7].

Machine learning models can be classified into individual models and hybrid ensemble machine learning. Various researchers have designed and implemented numerous prediction models using variations of classical machine learning methods and data mining strategies. Calisir and Dogantekin [8] proposed the LDA–MWSVM diabetes diagnosis technique. The system extracts and reduces features using Linear Discriminant Analysis (LDA) before classifying them using the Morlet Wavelet Support Vector Machine (MWSVM) classifier. Georga et al. [9] predicted short-term subcutaneous glucose concentrations using random forest decision trees and data from type 1 diabetes patients. Sardarinia et al. [10] created a logistic model for predicting the probability of developing type 2 diabetes, indicating that wrapper models, or models that look at a subset of features, can help clinical prediction models perform better. Iyer [11] investigated hidden patterns in a diabetic dataset using Naive Bayes and Decision Trees. Butwall and Kumar [12] developed a model for diabetic behaviour prediction that used a Random Forest Classifier. To predict the likelihood of diabetes, Sisodia & Sisodia [13] used Decision Tree, Support Vector Machine, and Naïve Bayes Classifiers. They demonstrated that Naïve Bayes performed better with an AUC of 0.819 than other models. Oleiwi et al. [14] proposed a machine learning algorithm utilizing significant features from the dataset and delivering findings similar to clinical outcomes. They trained Random Forest, Multi-layer Perceptron, and Radial Basis Function Network to find the best diabetes classifier. Bukhari et al. [15] used the Pima diabetes dataset to predict diabetes using an Artificial Neural Networks model with a configurable number of neurons in the hidden layers.

On the other hand, ensemble machine learning combines multiple machine learning models into a single predictive model. This learning technique offers dependable methods for dealing with uncertainty in the most difficult industrial challenges [16]. There are two types of ensemble methods: homogeneous and heterogeneous [17]. Homogeneous ensembles are models that use the same classification technique but have different parameter values. Boosting and bagging [18] fall into this category. AdaBoost, XGBoost, Additional Trees, and Gradient Booster are boosting techniques, whereas Random Forest is a bagging technique. Heterogeneous ensembles comprise models derived from multiple classification algorithms [19]. Stacking is a type of heterogeneous ensemble approach. In 2009, Sill et al. [20] developed a Feature-Weighted Linear Stacking (FWLS) method of blending linear regression and model trees. Compared to the standard linear stacking method, significant improvements in accuracy are achieved on the Netflix prize collaborative filtering dataset. Rider and Chawla [21] devised a method for gaining access to valuable information while maintaining patient privacy to find illness correlations, evaluate disease risk, and enhance it using ensemble learning.

Araújo et al. [22] modeled the behaviour of the preauthorization process using ensemble methods for the dental dataset. They merged the best classifiers using ensemble methods and random tree forest, Naive Bayes, SVM, and KNN classifiers. As meta-classifiers, Elkomy et al. [23] use Entropy, Gini index, wrapper-based feature selection, and SVM and KNN models to study illness progression prediction using genetic sequence data. Rajaraman et al. [16] used a combination of stacking and convolutional neural networks to predict tuberculosis from chest radio-graph images [16].

Although numerous machine learning approaches have been published in recent years, finding a machine learning approach that can more accurately predict diabetes remains a challenge. This research compares the performance of various classical machine learning models with the hybrid machine learning approach comprised of homogeneous and stacked generalization approaches. In addition, a method for constructing the stacking approach is presented in this study.

3- Research Methodology

3-1- Data Analysis

The Pima diabetes [24] and Early-stage-risk -prediction dataset [25] from Kaggle are used in this study. Pima diabetes dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. Other researchers have used this dataset and gained helpful insights, including Verma et al. [26] and Kumari et al. [27]. This dataset has 768 observations with 268 diabetes and 500 non-diabetes, including eight medical predictor attributes and one target attribute, as shown in Table 1. The eight predictor variables are numerical, and the target attribute is a binary value.

Table 1. Features of the Pima dataset

Attributes	Description of Attributes
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration after 2 hours in an oral glucose tolerance test. The glucose tolerance test can be used to screen for type 2 diabetes
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skinfold thickness (mm). Skinfold thickness can be used to estimate obesity and body fat distribution
Insulin	2-Hour serum insulin (mu U/ml). To determine whether there is a defect in islet function and metabolic disorder which are related to diabetes
BMI	Body mass index (Weight/Height, unit in kg/m^2. Body mass index (BMI) is a measure of obesity
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Outcome	Target variable

Early-stage diabetes dataset was collected from Sylhet Hospital in Bangladesh and authorized by the physician. There are 520 observations, including sixteen medical predictor attributes and one target attribute, as shown in Table 2. 320 of the 520 patients in the observations are positive, while the remaining 200 are negative.

Table 2. Features of the early-stage diabetes dataset

Attributes	Description of Attributes
Age	Age (years)
Gender	Gender of patient
Polyuria	Excessive urination
Polydipsia	Extreme thirst
Sudden Weight Loss	Loss of body weight within a very short amount of time
Weakness	Body fatigue
Polyphagia	A disorder that causes extreme hunger
Genital Thrush	A type of fungal infection inside the bowel or genitals
Visual Blurring	Distance vision blurring
Itching	Itchiness of the skin around the body
Irritability	An emotion that causes the person to feel irritated or agitated
Delayed Healing	A condition that prolongs the time taken for the body to heal wounds
Partial Paresis	The condition that weakens the muscles
Muscle Stiffness	Stiffness of the muscles
Alopecia	Hair loss condition
Obesity	Presence of excessive amounts of body fat
Class	Target variable

The mean, standard deviation, minimum, first quartile (25%), median (50%), third quartile (75%), and maximum are calculated using fundamental statistics, as illustrated in Tables 3 and 4. Based on this analysis, many attributes have a minimum value of 0. After investigation, the following attributes: Glucose, Blood Pressure, Skin Thickness Insulin and BMI have zero (0) values, which are invalid (impossible). Therefore we concluded that these zero values are missing values. Figure 1 shows the distribution of each missing attribute. Based on the distribution, Glucose and Blood Pressure have a normal distribution. Therefore, the mean of existing values replaces missing values for these values. Median is used to replace missing values in Skin Thickness, Insulin, and BMI. There is no missing for the early-stage diabetes dataset.

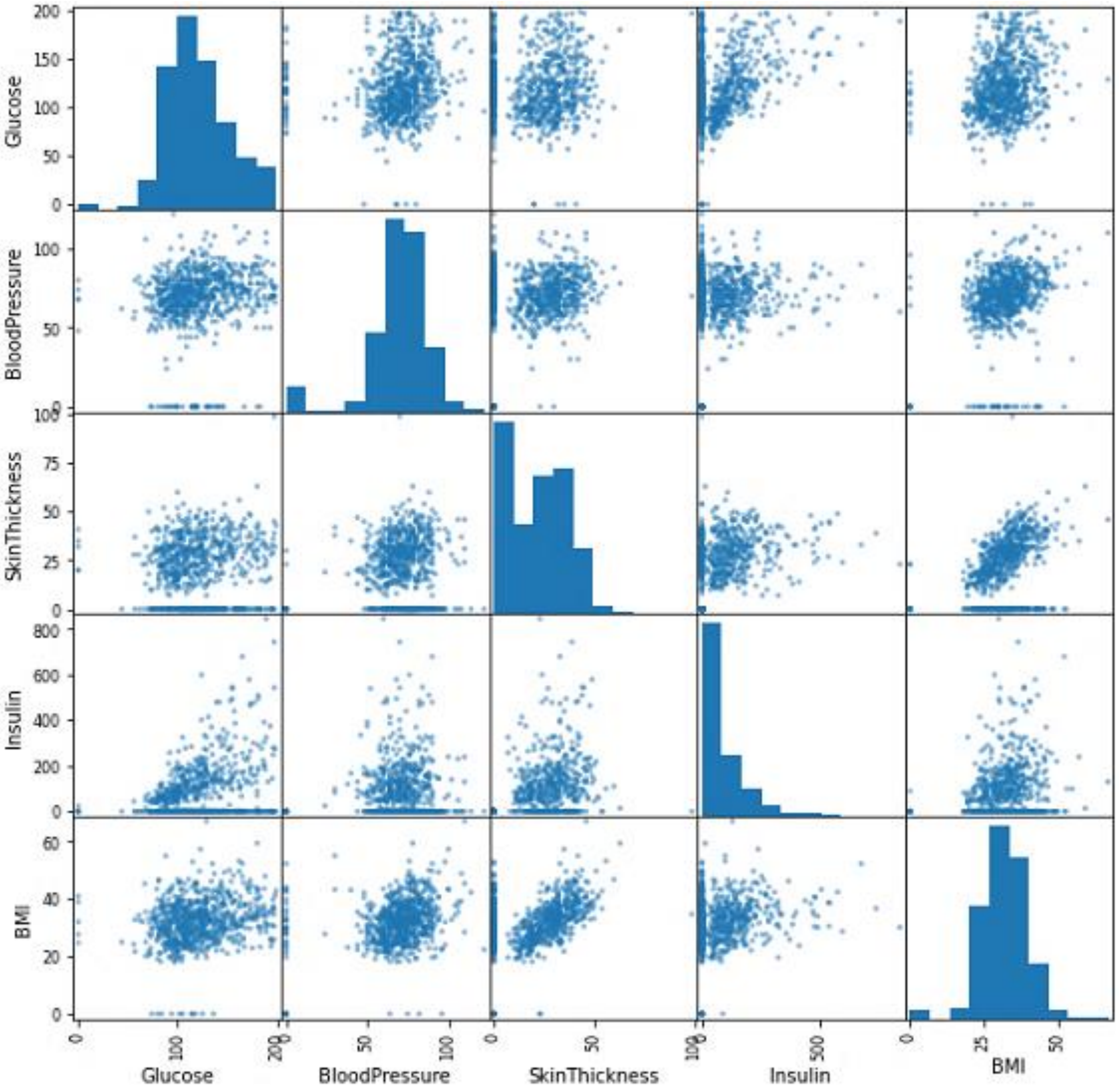


Figure 1. Distribution of missing attributes for the Pima diabetes dataset

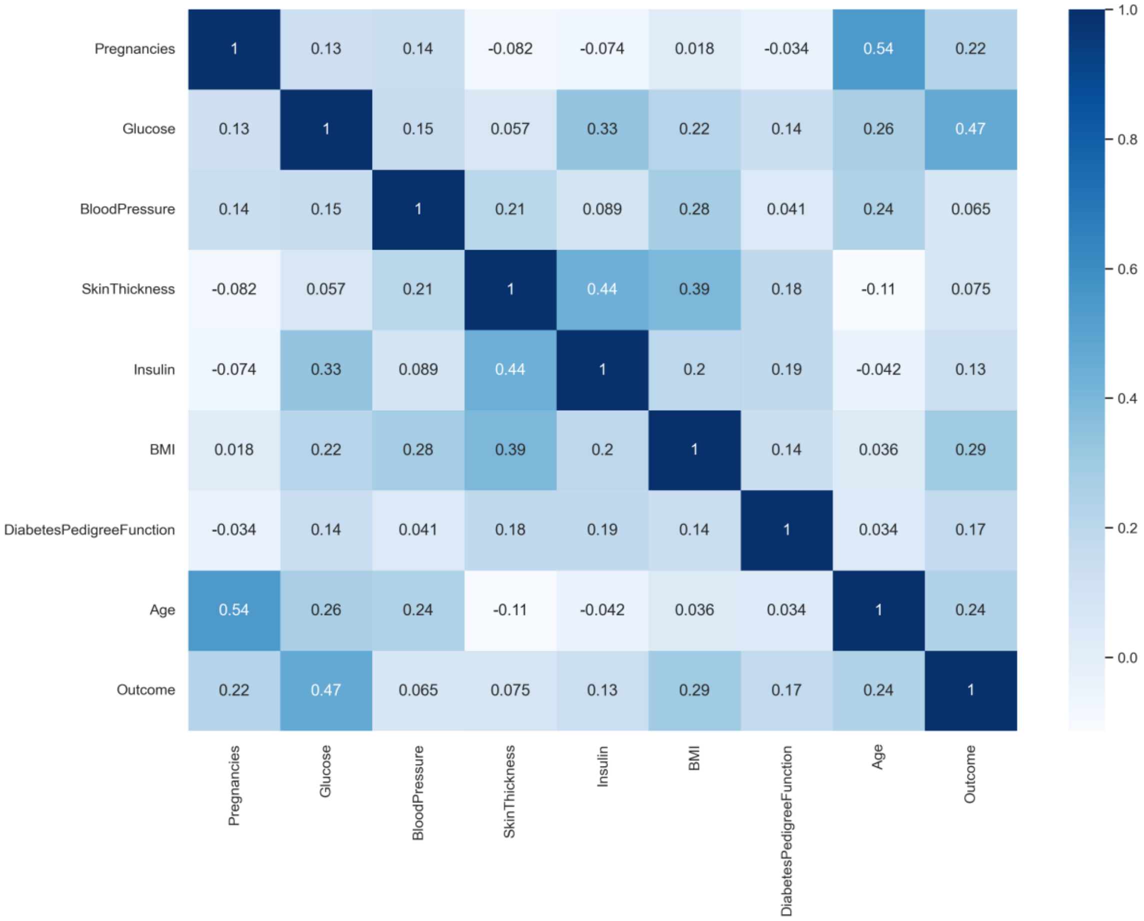
Table 3. The descriptive statistics of the Pima diabetes dataset

	Count	Mean	Std	Min	25%	50%	75%	Max
Pregnancies	768	3.8	3.4	0.0	1.0	3.0	6.0	17.0
Glucose	768	120.9	32.0	0.0	99.0	117.0	140.2	199.0
BloodPressure	768	69.1	19.4	0.0	62.0	72.0	80.0	122.0
SkinThickness	768	20.5	16.0	0.0	0.0	23.0	32.0	99.0
Insulin	768	79.8	115.2	0.0	0.0	30.5	127.2	846.0
BMI	768	32.0	7.9	0.0	27.3	32.0	36.6	67.1
DiabetesPedigreeFunction	768	0.5	0.3	0.1	0.2	0.4	0.6	2.4
Age	768	33.2	11.8	21.0	24.0	29.0	41.0	81.0
Outcome	768	0.3	0.5	0.0	0.0	0.0	1.0	1.0

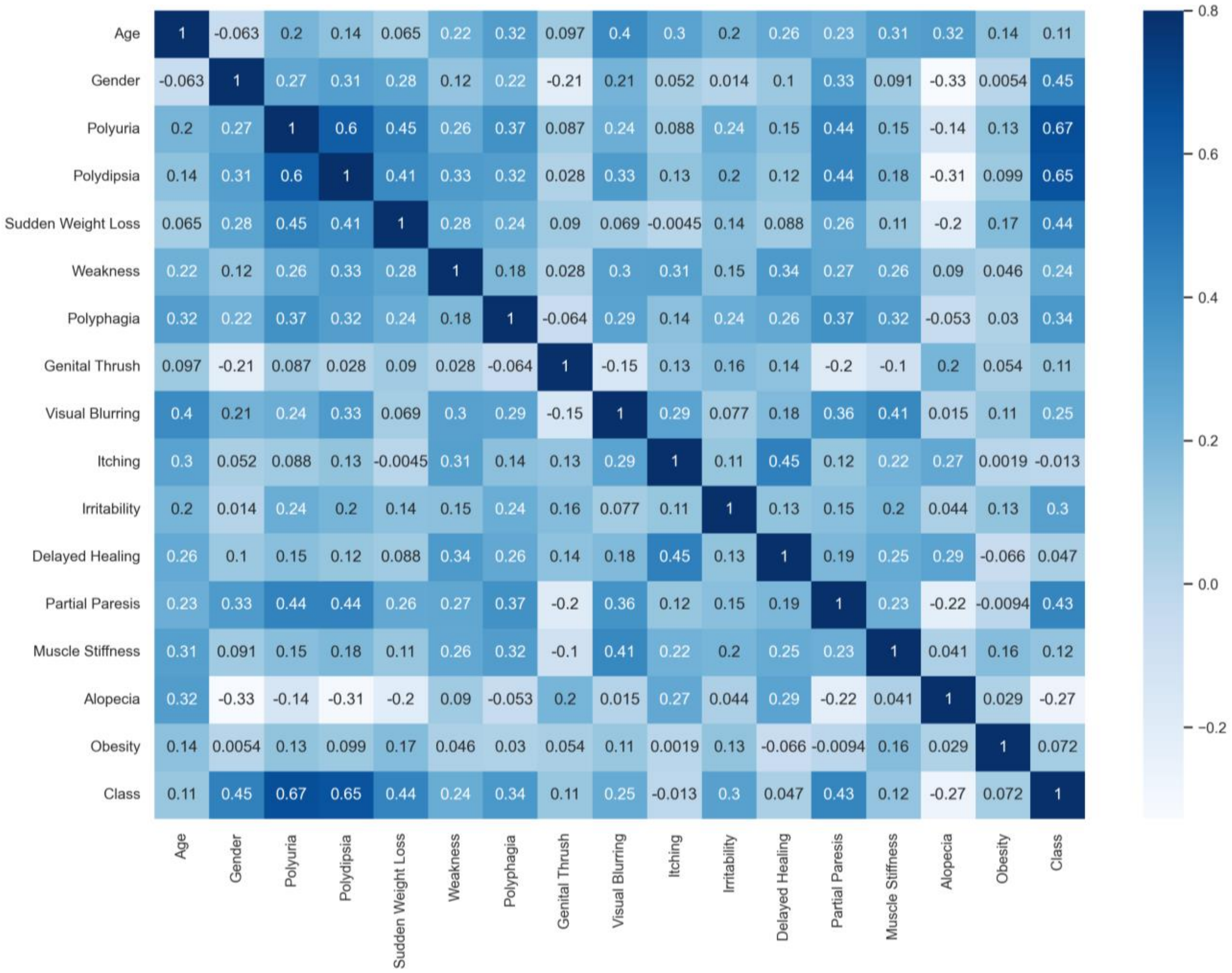
Table 4. The descriptive statistics of the early-stage diabetes dataset

	Count	Mean	Std	Min	25%	50%	75%	Max
Age	520	48.0	12.2	16.0	39.0	47.5	57.0	90.0
Gender	520	0.4	0.5	0.0	0.0	0.0	1.0	1.0
Polyuria	520	0.5	0.5	0.0	0.0	0.0	1.0	1.0
Polydipsia	520	0.4	0.5	0.0	0.0	0.0	1.0	1.0
Sudden Weight Loss	520	0.4	0.5	0.0	0.0	0.0	1.0	1.0
Weakness	520	0.6	0.5	0.0	0.0	1.0	1.0	1.0
Polyphagia	520	0.5	0.5	0.0	0.0	0.0	1.0	1.0
Genital Thrush	520	0.2	0.4	0.0	0.0	0.0	0.0	1.0
Visual Blurring	520	0.4	0.5	0.0	0.0	0.0	1.0	1.0
Itching	520	0.5	0.5	0.0	0.0	0.0	1.0	1.0
Irritability	520	0.2	0.4	0.0	0.0	0.0	0.0	1.0
Delayed Healing	520	0.5	0.5	0.0	0.0	0.0	1.0	1.0
Partial Paresis	520	0.4	0.5	0.0	0.0	0.0	1.0	1.0
Muscle Stiffness	520	0.4	0.5	0.0	0.0	0.0	1.0	1.0
Alopecia	520	0.3	0.5	0.0	0.0	0.0	1.0	1.0
Obesity	520	0.2	0.4	0.0	0.0	0.0	0.0	1.0
Class	520	0.6	0.5	0.0	0.0	1.0	1.0	1.0

Pearson’s correlation coefficient is used to find the relationship between attributes, as shown in Figures 2-a and 2-b. A significant correlation between pregnancy and age can be observed with a coefficient of 0.54. These attributes were clustered using K-means. After implementing this clustering, each record was assigned a class label (0 or 1). A similar correlation was carried out for the early-stage diabetes dataset. It was observed that Polydipsia has a high correlation with Polyuria, with a coefficient of 0.6.



(a)



(b)

Figure 2. a: Correlation Between Attributes of Pima diabetes dataset, b: Correlation between Attributes of Early-stage diabetes dataset

3-2- Data Partitioning

The prediction system was constructed using a five-fold cross-validation method on the training set [28]. Based on Anbananthen et al. [29], the entire data set is split into two parts: 70% of the data set is used to train the model, whereas 30% is used to test it. The data was divided into five equal portions at random. The models were trained on four subsets, with one subset left for validation. Each of the five subgroups was used exactly once to evaluate the performance. Figure 3 shows how 5-fold cross-validation works.

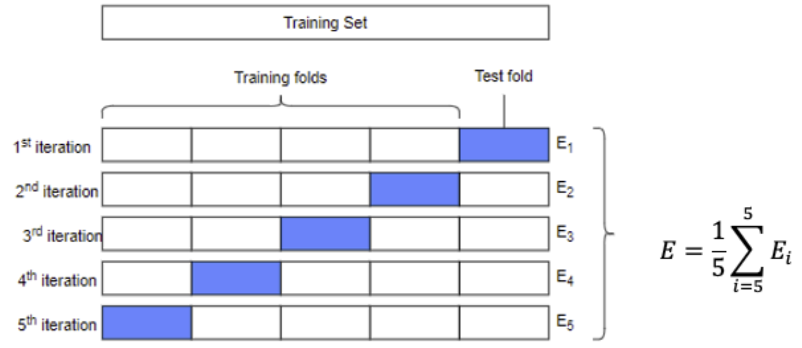


Figure 3. Train Test Split Details

3-3- Machine Learning Methods

The machine learning models proposed in this article are based on widely used techniques in classical and hybrid approach research [27, 30, 31]. All of these classifiers use the same training and test data. A total of four classical and five hybrid machine learning classifiers are employed in this study.

3-3-1- Classical Models

A Multi-layer Perceptron (MLP) is a multi-layer perceptron. It has three layers: input, output, and hidden. Each neuron computes an activation function. The input layer is responsible for introducing input values into the network; it does not contain any activation functions or other processing. The classification is done via the hidden layer. The output layer functions as a hidden layer and presents the result.

The k-Nearest Neighbour (KNN) classifier classifies new data points based on their distances between the clustered data points in its training data. A new data point will be assigned to the class with the closest data points from the new data point.

A support vector machine (SVM) is a supervised machine learning method that finds a hyperplane in N-dimensional space to identify the classes of data points in a dataset (N being the number of features).

A Naive Bayes (NB) classifier is a Bayes theorem-based probabilistic machine learning model. The Bayes theorem is a method for determining the conditional probability of events.

3-3-2- Ensemble Model

The Random Forest (RF) method is a form of supervised learning in which many decision trees with identical nodes are constructed. It generates numerous decision trees and merges their outputs to get an average solution.

AdaBoost is a supervised machine learning method that consists of an ensemble of stumps (one-level decision trees). Each of these stumps will vote on predicting a new data point. However, not all stumps have the same weight; therefore, the votes of certain stumps matter more than other stumps in the ensemble.

XGBoost is a supervised ensemble machine learning technique based on the decision tree method. It uses a gradient-boosting framework. Each treemap an input data point to one of its leaves containing a continuous score. It minimizes a regularised objective function. It seeks to reduce a regularised objective function, which is a combination of a convex loss function and a model complexity penalty term.

Extra Trees is a supervised machine learning approach that constructs many unpruned decision trees from a training dataset. Predictions are made by averaging the decision trees' predictions or by a majority vote.

The Gradient Boosted Tree (GBT) regression tree model is one of the most successful machine learning techniques for predictive investigations. Gradient boosting comprises an optimized loss function, a weak learner that generates predictions, and an additive model that combines weak learners to minimize the loss function. Each decision tree iteration improves the output value by adjusting the input variable's weights or bias coefficients.

3-3- 3- Stacked Generalization

A stacked generalization was built by combining multiple classifiers in a two-tier structure, as depicted in Figure 4. This structure was used for Pima and Early-stage diabetes datasets. The first layer contained the base/individual models [32]. It involves training several classifiers as base learners in the first layer. The second layer had a single classifier referred to as the meta-classifier (learner). The output of the base learners' in the first layer is used as an input to retrain the second layer classifier. Not all the output from the base learner will be used as input for the meta-classifier. Only an optimal combination of the base-level classifier will be selected. In this study, Tabu Search (TS) algorithm [33, 34] is utilized to determine the optimal combination of the base-level classifier. The meta-classifier allocates weights from each base-level classifier based on their cross-validated prediction performance. All four classical and five hybrid models discussed above are used as the base classifier in this research. The base classifiers are trained with the k-1 fold of the training data. Each training fold is used to train base classifiers, and predictions are derived from the validation fold. Predicted labels from each cross-validation testing fold of the base classifiers are combined and applied to train the meta-classifier as described in the algorithm in Figure 5 [35, 36]. This layer is tested using the test data.

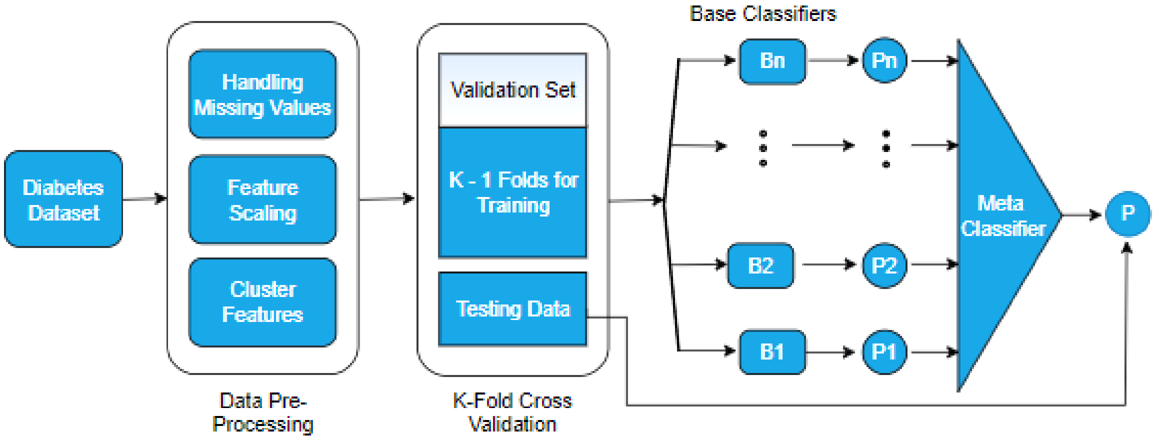
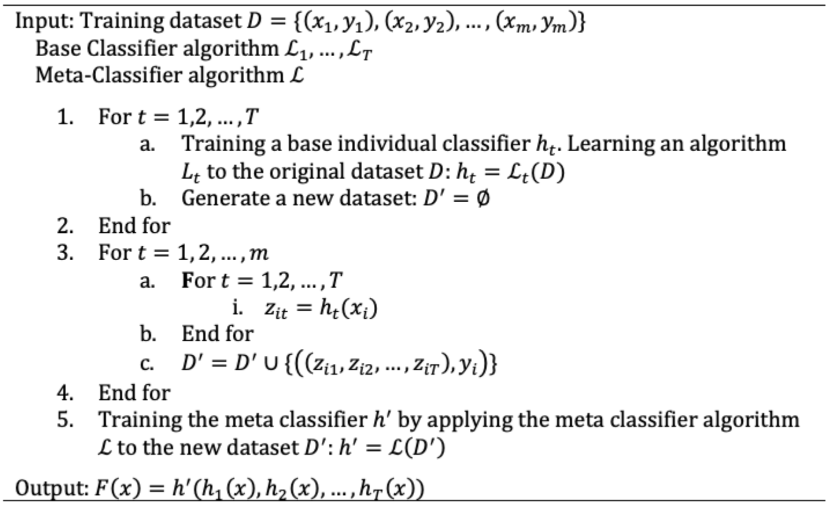


Figure 2. The architecture of the proposed stacked generation model



```
Input: Training dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$
  Base Classifier algorithm $\mathcal{L}_1, \dots, \mathcal{L}_T$
  Meta-Classifier algorithm $\mathcal{L}$

    1. For $t = 1,2, \dots, T$
         a. Training a base individual classifier $h_t$. Learning an algorithm
            $L_t$ to the original dataset $D$: $h_t = \mathcal{L}_t(D)$
         b. Generate a new dataset: $D' = \emptyset$
    2. End for
    3. For $t = 1, 2, \dots, m$
         a. For $t = 1,2, \dots, T$
              i. $z_{it} = h_t(x_i)$
         b. End for
         c. $D' = D' \cup \{((z_{i1}, z_{i2}, \dots, z_{iT}), y_i)\}$
    4. End for
    5. Training the meta classifier $h'$ by applying the meta classifier algorithm
       $\mathcal{L}$ to the new dataset $D'$: $h' = \mathcal{L}(D')$

Output: $F(x) = h'(h_1(x), h_2(x), \dots, h_T(x))$
```

Figure 3. The algorithm for the proposed approach

Based on the accuracy of the individual model, GBT is chosen as a meta-classifier, while all the nine classifiers serve as base-level classifiers. Adaboost, XGBoost, and GBT are chosen as the optimal combination to create the stacked generalization model based on the TS algorithm. B1, B2,... Bn in Figure 4 depicts the base classifier. The prediction models P1, P2,...Pn is generated based on the training model and test data.

The flowchart (Figure 6) illustrates the methodology adopted for this research. First, the data set was collected and preprocessed. The preprocessing steps include filling in missing values, feature combination, and dimensionality reduction. After preprocessing, the data is divided into training and testing groups. All models were trained using 5-fold cross-validation.

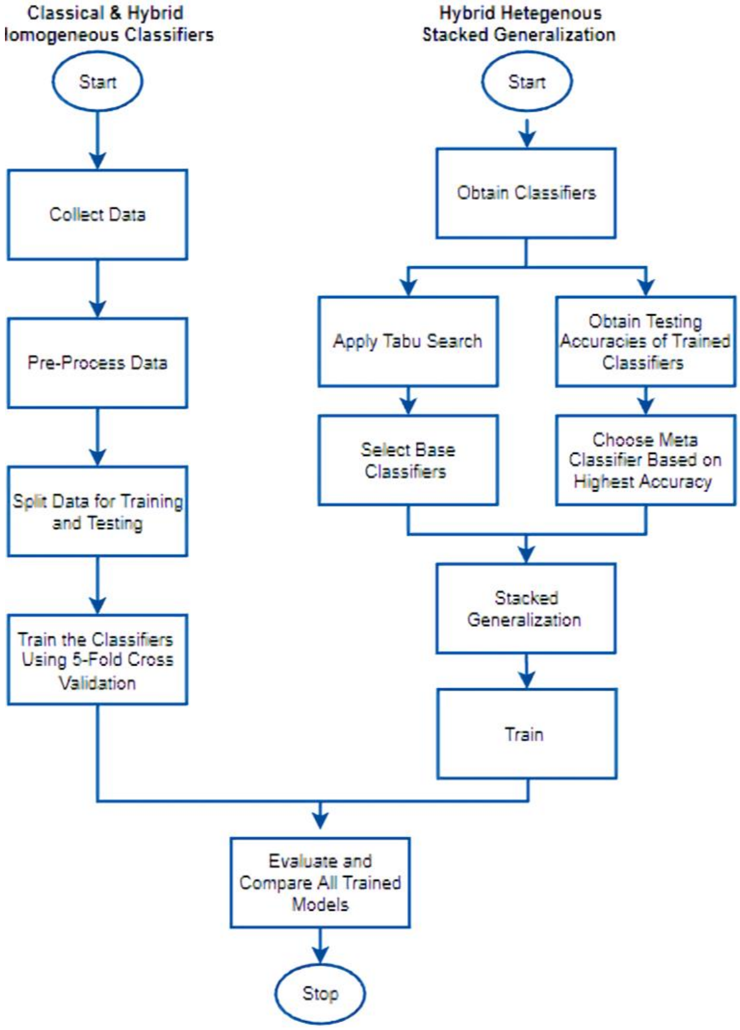

Classical & Hybrid
Iomogeneous Classifiers
Hybrid Hetegenous
Stacked Generalization
Start
Start
Collect Data
Obtain Classifiers
Pre-Process Data
Apply Tabu Search
Obtain Testing Accuracies of Trained Classifiers
Select Base Classifiers
Choose Meta Classifier Based on Highest Accuracy
Split Data for Training and Testing
Stacked Generalization
Train the Classifiers Using 5-Fold Cross Validation
Train
Evaluate and Compare All Trained Models
Stop

Figure 4. Flowchart of the research methodology

First, to develop the stacked generation model, find the optimal base learner based on the Tabu search algorithm. The combination of Adaboost, XGBoost, and GBT is chosen as the optimal base learners for the stack. At the same time, based on the accuracy of the classical and homogeneous model, select the meta-classifier. GBT, which has the highest accuracy, is chosen as the meta-classifier in this research. Stacking was built by combining the base and meta-learners. In stacking, the algorithm learns (trains) by using the combined outputs of base models as input for the meta-classifier to generate the output prediction of the stacking.

- Level 0: Adaboost, XGBoost, and GBT learn to make predictions from the training data.
- Level 1: The output of Adaboost, XGBoost, and GBT is used as input for the meta-classifier (GBT). GBT learns to make predictions from this data.

4- Results and Discussion

This section compares and analyzes the machine learning approaches on Pima and the early-stage dataset. The performance of the diabetes prediction was assessed using the accuracy, precision, recall, F1 score, Area Under the Curve (AUC), and Matthews Correlation Coefficient (MCC), as shown in Table 5. These metrics compare the hybrid homogeneous models, Stacked Generalization, and the classical model. Accuracy is a widely used statistic for assessing a model’s predictive capability and accuracy. The Confusion Matrix (Figure 7) gives us a matrix as an output that describes the model’s overall performance.

Table 5. List of evaluation measures

Metric	Definition	Formula
Accuracy (ACC)	The ratio of correct predictions to total input samples.	$ACC = \frac{TP + TN}{N}$
Precision (P)	The ratio of correct positive results to positive results is predicted by the classifier.	$P = \frac{TP}{TP + FP}$
Recall (R)	The number of correctly identified positive findings is divided by the total number of relevant samples.	$R = \frac{TP}{TP + FN}$
F1-Score (F1)	Used to determine the correctness of a test. It is defined as the harmonic mean of precision and recall, with a range of [0,1]. It indicates both the precision and robustness of the classifier. F1 seeks a combination of precision and recall.	$F1 = 2 \times \frac{P \times R}{P + R}$
Specificity (S)	The ratio between correctly identified as negative to actually negative.	$S = \frac{TN}{TN + FP}$
Area Under Curve (AUC)	The receiver operating characteristics (ROC) curve summarises the receiver operating characteristics, which indicates a classifier’s ability to discriminate classes. The AUC value represents the model’s ability to discriminate between positive and negative classes; the higher the AUC value, the better.	-
Matthews Correlation Coefficient (MCC)	A statistic that indicates the correlation between true classes and predicted labels	$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP).(TP + FN).(TN + FP).(TN + FN)}}$

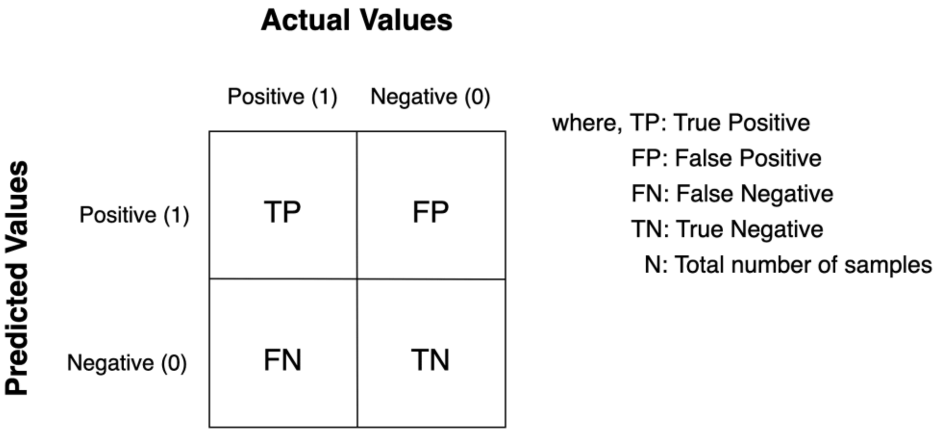


Figure 5. Confusion matrix for predictions

Tables 6 and 7 compare the performance of four classical models (MLP, kNN, SVM, NB), and six hybrid models (RF, AdaBoost, XGB, Extra Trees, GBT, Stacked Generalization) on the Pima and Early-stage diabetes dataset. The hybrid models outperform the classical models on both datasets in terms of prediction accuracy, F1 Score, and Matthews Correlation Coefficient, as seen in both tables. This indicates hybrid models generalize better. The performance of all hybrid models is relatively comparable with the Stacked Generalization model achieving the highest accuracy for PIMA and Early-stage, respectively, of 83.9% and 98.5%. Tables 6 and 7 show that the KNN performs poorly, with 72.4% accuracy for Pima and 86.9% for the early-stage data set.

Table 6. Performance of various base learners models on Pima diabetes dataset

	Accuracy	Precision	Recall	F1-Score	Specificity	AUC	MCC
MLP	0.750	0.821	0.438	0.571	0.941	0.860	0.458
KNN	0.724	0.738	0.425	0.539	0.908	0.809	0.390
SVM	0.771	0.830	0.493	0.621	0.941	0.717	0.506
Naïve Bayes	0.750	0.745	0.521	0.613	0.891	0.857	0.452
RF	0.781	0.804	0.562	0.661	0.916	0.739	0.525
Ada Boost	0.792	0.690	0.753	0.750	0.773	0.860	0.580
XG Boost	0.792	0.690	0.753	0.750	0.773	0.861	0.580
Extra Trees	0.807	0.833	0.616	0.709	0.924	0.814	0.584
GBT	0.812	0.723	0.753	0.769	0.807	0.870	0.616
Stacked Generalization	0.839	0.837	0.767	0.783	0.899	0.877	0.655

Table 7. Performance of various base learners on Early-stage dataset

	Accuracy	Precision	Recall	F1-Score	Specificity	AUC	MCC
MLP	0.923	0.950	0.927	0.938	0.917	0.922	0.837
KNN	0.869	0.933	0.854	0.892	0.896	0.875	0.732
SVM	0.946	0.952	0.963	0.957	0.917	0.940	0.884
Naïve Bayes	0.923	0.974	0.902	0.937	0.958	0.930	0.843
RF	0.938	0.974	0.927	0.950	0.958	0.943	0.872
Ada Boost	0.969	0.988	0.963	0.975	0.979	0.971	0.935
XG Boost	0.954	0.975	0.951	0.963	0.958	0.955	0.902
Extra Trees	0.962	0.975	0.963	0.969	0.958	0.961	0.918
GBT	0.977	0.976	0.988	0.982	0.958	0.973	0.950
Stacked Generalization	0.985	0.988	0.988	0.988	0.979	0.983	0.967

In terms of precision, recall, F1 score, AUC value, and MCC value, the Stacked Generalization Model also achieves the highest score on both datasets. This indicates that compared to the other models, Stacked Generalization can correctly identify 83.1% (Pima) and 98.8% (Early-stage) individuals as having diabetes out of those who truly have it. It is important that we don't start treating a person who doesn't have diabetes, but the model predicted it as having diabetes. Recall for the Stacked Generalization model is 76.7 (Pima) and 98.8% (Early-stage). Recall measures the precision with which the model can identify the relevant details. What if a person with diabetes does not receive treatment because the model indicates they do not have diabetes. In terms of AUC, Stacked Generalization will be able to distinguish the patients with diabetes and those who don't 98.3% of the time. In terms of specificity, the most classical model outperforms the hybrid model. It measures how many patients do not have diabetes among all the people who actually don't have diabetes.

In terms of the specificity score, most classical classifiers perform slightly better than hybrid machine learning models. Sometimes specifity value alone cannot decide the performance of the model, and it has to be looked at along with sensitivity values. Therefore, in order to check if the AUC value of the proposed approach is better than another classifier, we have plotted the ROC-AUC curve (Figures 8-a and 8-b). Based on the AUC values of the curve, we can observe that the AUC value of the hybrid method, especially Stacked Generalization, performs better than all machine learning models.

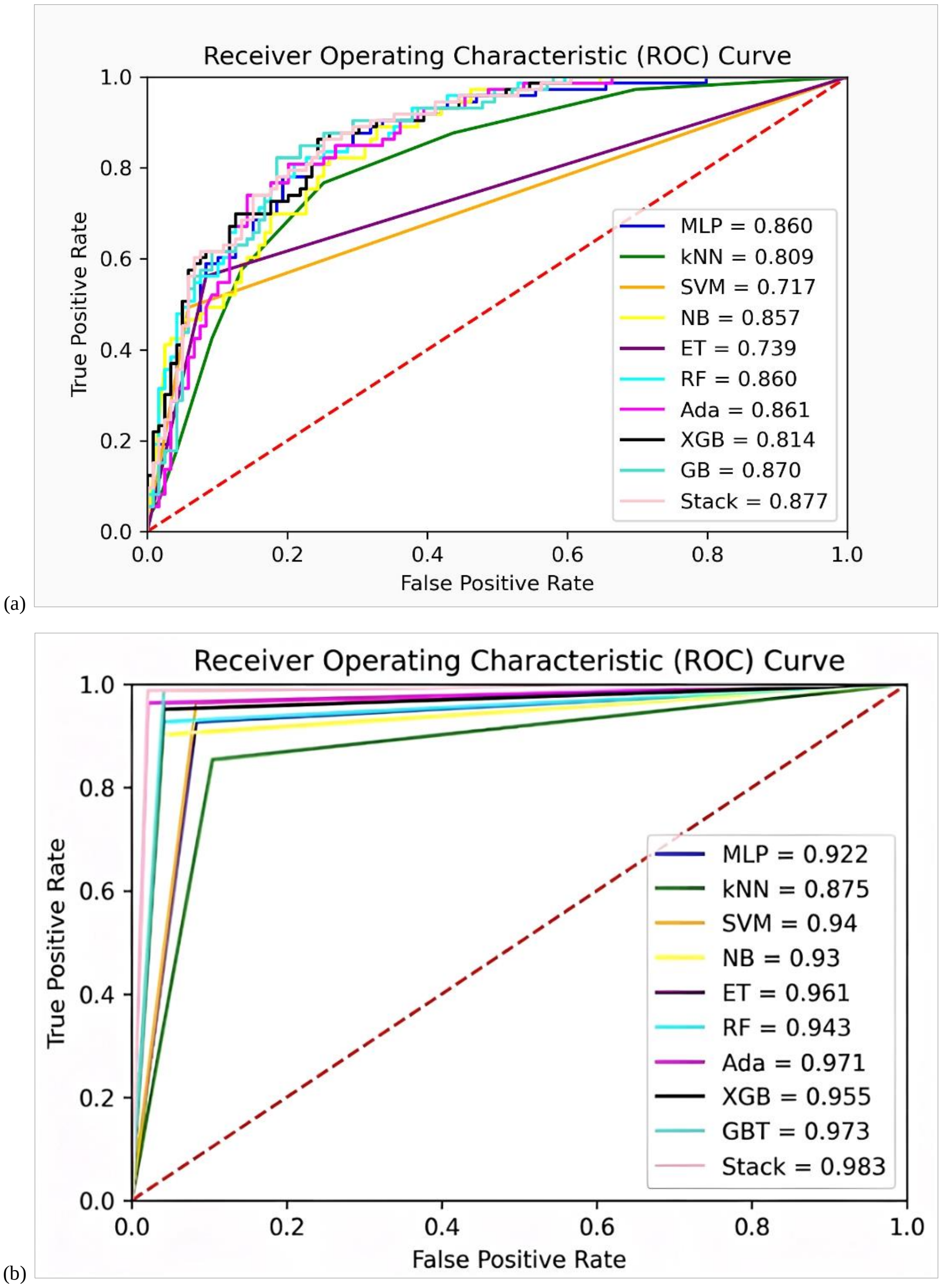


Figure 6. a) Comparison of ROC-AUC scores of Pima diabetes dataset, b) Comparison of ROC-AUC scores of Early-stage diabetes dataset

4-1- Discussion

In this research, we evaluate the performance of several classical machine learning models with the hybrid machine learning approach to find a machine learning approach that can more accurately predict diabetes. The hybrid learning method provides more accurate predictions and outperforms any classical model. It was observed that the spread or dispersion of the predictions is reduced by using hybrid learning in these two datasets, therefore increasing the accuracy of the models. Each classical model is able to learn only one aspect of data structure, and with the hybrid approach, each model accurately captures a different aspect of data structure. Experimental results show (Tables 6 and 7) that hybrid learning is better overall than the classical method. Heterogeneous Stack generalized models outperform homogenous hybrid learning algorithms. Combining the strengths of different base model types compensates for the homogeneous base model's variances and biases in this problem. Additionally, certain models excel at modeling a particular data feature, whilst others excel at modeling a different aspect, and predictions come from combining the strengths of different models.

Table 8 compares the current state of the art to the best model from our analysis, the Stacked Generalized model. Mirshahvalad et al. [37] used an ensemble boosting algorithm with a perceptron algorithm. They achieved an accuracy of 79%, whereas Sisodia & Sisodia [13] used modified machine learning algorithms with efficient coding to attain 76.30% accuracy. Kumari et al. [27] used a more accurate Ensemble Soft Voting classifier that combines Logistic Regression, Naive Bayes, and Random Forest to predict diabetes, achieving an accuracy of 79.1%. Based on the results displayed in Table 8, it can be seen that the Stacked Generalized Models hybrid model with an accuracy of 83.9% outperformed the models proposed by the previous works. This can be attributed to our proposed solution's base and meta-classifier.

Table 8. Comparative analysis

Work	Model	Accuracy
Mirshahvalad & Zanjani (2017) [37]	Ensemble Perceptron Algorithm	74.0%
Sisodia & Sisodia (2018) [13]	Modified Machine Learning Algorithms	76.3.0%
Kumari et al. (2021) [27]	Ensemble Soft Voting Classifier	79.1%
Proposed work	Stacked Generalization	83.9%

This work does have certain limitations. The early-stage dataset has more attributes than the Pima dataset, and the sample size and the number of features in the Pima dataset are small. This dataset does not consider other diabetes risk factors, such as lifestyle, eating habits, smoking, genetics, and stress.

5- Conclusion

Diabetes mellitus is a disease that is increasingly prevalent in people nowadays. As a result, early detection of this disease is critical. The main aim of this paper is to provide insight into the preferable diabetes prediction model. A comparative evaluation of four classical models, i.e., MLP, KNN, SVM, NB, with five homogenous hybrid models, i.e., RF, Ada Boost, XGBoost, Extra Trees, and GBT, with the heterogeneous hybrid model using Stacked Generalization is performed. Additionally, this paper provides an approach for constructing stacked generalization. TS algorithm is used to select the optimal combination of base classifiers. This algorithm selects Adaboost, XGBoost, and GBT as the optimal combinations to produce the base learner model, with GBT used as a meta-classifier. Experiments with diabetes data revealed that the heterogeneous stacked generalization model outperforms classical and hybrid homogenous models, achieving an accuracy of 83.9% for Pima and 98.5% for early-stage datasets. The study's findings may assist clinicians in diagnosing diabetes and making more educated clinical management decisions. While our research has demonstrated that the stacked generalization approach is a good option for predicting diabetes, it does have certain limitations. First, our conclusion is based on the minimal facts available. A bigger dataset and more features, including risk factors, lifestyle, and real-time diabetes data, will be collected in the future to complete the research findings. Furthermore, more research is needed to identify the best meta and base learner combination. Deep learning techniques like the Deep Convolution Neural Network (DCNN) should be further explored to see if they can offer a better and more accurate solution to this problem.

6- Declarations

6-1- Author Contributions

Conceptualization, K.S.M.A. and M.M.A.B.; methodology, K.S.M.A. and M.M.A.B.; validation, R.K., and S.K.; formal analysis, R.K.; investigation, S.K.; writing—original draft preparation, K.S.M.A.; writing—review and editing, R.K., and S.K.; visualization, M.M.A.B.; supervision, K.S.M.A.; project administration, S.K. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

Publicly available datasets were analyzed in this study. This data can be found here: PIMA diabetes Dataset: *https://www.kaggle.com/uciml/pima-indians-diabetes-database.* Early-stage diabetes Dataset: *https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset.*

6-3- Funding

This study is supported by the Faculty of Science and Technology (FIST), Multimedia University, Malaysia.

6-4- Institutional Review Board Statement

Not applicable.

6-5- Informed Consent Statement

Not applicable.

6-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies, have been completely observed by the authors.

7- References

[1] Wild, S., Roglic, G., Green, A., Sicree, R., & King, H. (2004). Global Prevalence of Diabetes: Estimates for the year 2000 and projections for 2030. Diabetes Care, 27(5), 1047–1053. doi:10.2337/diacare.27.5.1047.

[2] International Diabetes Federation (IDF). (2021). Learning of diabetes facts figures. International Diabetes Federation, Brussels, Belgium. Available online: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html (accessed on August 2022).

[3] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A. A., Ogurtsova, K., Shaw, J. E., Bright, D., & Williams, R. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Research and Clinical Practice, 157, 1–10,. doi:10.1016/j.diabres.2019.107843.

[4] Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Research and Clinical Practice, 138, 271–281. doi:10.1016/j.diabres.2018.02.023.

[5] Seery, C. (2019). Diabetes Prevalence. The global diabetes community, Available online: https://www.diabetes.co.uk/diabetes-prevalence.html (accessed on August 2022).

[6] Syaifuddin, M., & Muthu Anbananthen, K. S. (2013). Framework: Diabetes management system. IMPACT-2013. doi:10.1109/mspct.2013.6782099.

[7] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104–116. doi:10.1016/j.csbj.2016.12.005.

[8] Çalişir, D., & Doğantekin, E. (2011). An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Systems with Applications, 38(7), 8311–8315. doi:10.1016/j.eswa.2011.01.017.

[9] Georga, E. I., Protopappas, V. C., Polyzos, D., & Fotiadis, D. I. (2012). A predictive model of subcutaneous glucose concentration in type 1 diabetes based on Random Forests. 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. doi:10.1109/embc.2012.6346567.

[10] Sardarinia, M., Akbarpour, S., Lotfaliany, M., Bagherzadeh-Khiabani, F., Bozorgmanesh, M., Sheikholeslami, F., Azizi, F., & Hadaegh, F. (2016). Risk factors for incidence of cardiovascular diseases and all-cause mortality in a middle-eastern population over a decade follow-up: Tehran lipid and glucose study. PLoS ONE, 11(12), 12. doi:10.1371/journal.pone.0167623.

[11] Iyer, A., S, J., & Sumbaly, R. (2015). Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process, 5(1), 01–14. doi:10.5121/ijdkp.2015.5101.

[12] Butwall, M., & Kumar, S. (2015). A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier. International Journal of Computer Applications, 120(8), 36–39. doi:10.5120/21249-4065.

[13] Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. Procedia Computer Science, 132, 1578–1585. doi:10.1016/j.procs.2018.05.122.

[14] Oleiwi, A., Shi, L., Tao, Y., & Wei, L. (2020). A comparative analysis and risk prediction of diabetes at early stage using machine learning approach. International Journal of Future Generation Communication and Networking, 13(3), 4151-4163.

[15] Bukhari, M. M., Alkhamees, B. F., Hussain, S., Gumaei, A., Assiri, A., & Ullah, S. S. (2021). An Improved Artificial Neural Network Model for Effective Diabetes Prediction. Complexity, 2021, 1–10. doi:10.1155/2021/5525271.

[16] Rajaraman, S., Candemir, S., Xue, Z., Alderson, P. O., Kohli, M., Abuya, J., Thoma, G. R., & Antani, S. (2018). A novel stacked generalization of models for improved TB detection in chest radiographs. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). doi:10.1109/embc.2018.8512337.

[17] Graczyk, M., Lasota, T., Trawiński, B., Trawiński, K. (2010). Comparison of Bagging, Boosting and Stacking Ensembles Applied to Real Estate Appraisal. Intelligent Information and Database Systems. ACIIDS 2010, Lecture Notes in Computer Science, 5991, Springer, Berlin, Germany. doi:10.1007/978-3-642-12101-2_35.

[18] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140. doi:10.1007/bf00058655.

[19] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. doi:10.1016/s0893-6080(05)80023-1.

[20] Sill, J., Takács, G., Mackey, L., & Lin, D. (2009). Feature-weighted linear stacking. arXiv preprint. doi:10.48550/arXiv.0911.0460

[21] Rider, A. K., & Chawla, N. V. (2013). An Ensemble Topic Model for Sharing Healthcare Data and Predicting Disease Risk. Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. doi:10.1145/2506583.2506640.

[22] Araújo, F. H. D., Santana, A. M., & de A. Santos Neto, P. (2016). Using machine learning to support healthcare professionals in making preauthorization decisions. International Journal of Medical Informatics, 94, 1–7. doi:10.1016/j.ijmedinf.2016.06.007.

[23] Elkomy, G., Sallam, E., & Elgokhy, S. (2017). A stacked generalization method for disease progression prediction. 2017 13th International Computer Engineering Conference (ICENCO). doi:10.1109/icenco.2017.8289772.

[24] Kaggle Inc. (2016). Pima Indians Diabetes Databases. Available online: https://www.kaggle.com/uciml/pima-indians-diabetes-database (accessed on August 2022).

[25] Dutta, I. (2020). Early-stage diabetes. Kaggle. Available online: https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset (Accessed on April 2022).

[26] Verma, D., & Mishra, N. (2017). Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques. 2017 International Conference on Intelligent Sustainable Systems (ICISS). doi:10.1109/iss1.2017.8389229.

[27] Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering, 2, 40–46. doi:10.1016/j.ijcce.2021.01.001.

[28] Cao, H., Peng, J., Zhou, Z., Sun, Y., Wang, Y., & Liang, Y. (2022). Insight into the defluorination ability of per-and polyfluoroalkyl substances based on machine learning and quantum chemical computations. Science of the Total Environment, 807, 151018. doi:10.1016/j.scitotenv.2021.151018.

[29] Anbananthen, K. S. M., Subbiah, S., Chelliah, D., Sivakumar, P., Somasundaram, V., Velshankar, K. H., & Khan, M. K. A. A. (2021). An intelligent decision support system for crop yield prediction using hybrid machine learning algorithms. F1000Research, 10. doi:10.12688/f1000research.73009.1.

[30] Xiao, H., Xiao, Z., & Wang, Y. (2016). Ensemble classification based on supervised clustering for credit scoring. Applied Soft Computing Journal, 43, 73–86. doi:10.1016/j.asoc.2016.02.022.

[31] Abdollahi, J., & Nouri-Moghaddam, B. (2022). Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction. Iran Journal of Computer Science, 5(3), 205–220. doi:10.1007/s42044-022-00100-1.

[32] Talukdar, S., Pal, S., & Singha, P. (2021). Proposing artificial intelligence based livelihood vulnerability index in river islands. Journal of Cleaner Production, 284, 124707. doi:10.1016/j.jclepro.2020.124707.

[33] Dhahri, H., Rahmany, I., Mahmood, A., Al Maghayreh, E., & Elkilani, W. (2020). Tabu Search and Machine-Learning Classification of Benign and Malignant Proliferative Breast Lesions. BioMed Research International, 2020. doi:10.1155/2020/4671349.

[34] Guo, B., Hu, J., Wu, W., Peng, Q., & Wu, F. (2019). The Tabu_genetic algorithm: A novel method for hyper-parameter optimization of learning algorithms. Electronics (Switzerland), 8(5), 1–19. doi:10.3390/electronics8050579.

[35] Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. Expert Systems with Applications, 38(1), 223–230. doi:10.1016/j.eswa.2010.06.048.

[36] Liao, Z., Su, M., Ning, G., Liu, Y., Wang, T., & Zhou, J. (2021). A Novel Stacked Generalization Ensemble-Based Hybrid PSVM-PMLP-MLR Model for Energy Consumption Prediction of Copper Foil Electrolytic Preparation. IEEE Access, 9, 5821–5831. doi:10.1109/ACCESS.2020.3048714.

[37] Mirshahvalad, R., & Zanjani, N. A. (2017). Diabetes prediction using ensemble perceptron algorithm. 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN), IEEE, Girne, Northern Cyprus, 17634291. doi:10.1109/cicn.2017.8319383.