Towards a Stacking Ensemble Model for Predicting Diabetes Mellitus using Combination of Machine Learning Techniques

Abdulaziz A Alzubaidi, Sami M Halawani, Mutasem Jarrah
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Diabetes Mellitus (DM) is a chronic disease affecting the world's population, it causes long-term issues such as kidney failure, blindness, and heart disease, hurting one's quality of life. Diagnosing diabetes mellitus in an early stage is a challenge and a decisive decision for medical experts, as delay in diagnosis leads to complications in controlling the progression of the disease. Therefore, this research aims to develop a novel stacking ensemble model to predict diabetes mellitus a combination of machine learning models, where an ensemble of Prediction classifiers was used, such as Random Forest (RF), Logistic Regression (LR), as base learners' models, and the Extreme gradient Boosting model (XGBoost) as a Meta-Learner model. The results indicated that our proposed stacking model can predict diabetes mellitus with 83% accuracy on Pima dataset and 97% with DPD dataset. In conclusion, our proposed model can be used to build a diagnostic application for diabetes mellitus, as recommend testing our model on a huge and diverse dataset to obtain more accurate results.**

***Keywords*—DM; Diabetes Mellitus; Stacking; Ensemble learning; Machine Learning; Random Forest (RF); Logistic Regression (LR); Extreme Gradient Boosting model (XGBoost)**

I. Introduction

Diabetes Mellitus mainly leads to chronic hyperglycemia considering low insulin quantities in the bloodstream [1]. Insulin plays an important role in glucose level lowering in the blood, carbohydrate anabolism, physical growth, cell reproduction, and protein and fat anabolic statute [2]. Therefore, severe difficulties are associated explicitly with Diabetes Mellitus concerning people's quality of life. The chronic diseases caused by Diabetes Mellitus have heart failure, kidney failure, blindness, and cardiovascular illnesses [3]. These conditions induce a high pile in mortality rates and pressures on personal life [4]. According to estimates, there was 463 million people worldwide with diabetes in 2019 [5]. Moreover, by 2030, that number is expected to rise to 578 million and then 700 million (2045).

Urban areas (10.8% percent) have a higher prevalence than rural areas (7.2% percent), and high-income countries (10.4% percent) have a higher prevalence than low-income ones (4.0% percent) [5]. 50.1% percent of people with diabetes don't know they have the disease.

According to estimates, there is 7.5% percent (374 million) people worldwide who have impaired glucose tolerance, and that number is expected to rise to 8.0% percent (454 million) by 2030 and 8.6% percent (548 million) by 2045 [5]. There are two prevalent forms of DM: Type-1-Diabetes (T1DM), an autoimmune syndrome that results in the death of beta cells in the pancreas that produce insulin, and Type-2-Diabetes (T2DM), which is a chronic condition that frequently results in abnormally high blood sugar levels (glucose) [6]. T1DM affects about 10% percent of patients under 30, whereas T2DM affects about 90% percent of diabetics over 30% percent [6]. Doctors use trial results from agreed-upon studies to distinguish between these types and then specify the best treatment options based on the form they have discovered. Medical experts occasionally disagree on the proper type of diagnosis, which makes treating the illness challenging [7].

Diabetes is becoming more prevalent worldwide, particularly in middle-income nations [8]. Therefore, we need to conduct this study to predict diabetes using machine learning methods to support doctors in providing the most suitable treatment strategy.

It was noted during the literature reviews that the emphasis is on ensemble machine learning techniques for predicting diabetes mellitus therapy and prevention are challenging due to suitable policies to provide environments that support healthy behaviors and a lack of quality health care in various settings.

The Sustainable Development Community seeks to eliminate premature mortality for various illnesses, including diabetes, by 2030 [8]. As a result, experts are continually researching multiple facets of DM where many machine learning techniques are used such as the RF model, which is a great choice in binary classification processes, for example in diabetes mellitus, the outcome is that the patient is diabetic or not diabetic, where the random forest depends on ensemble learning method (bagging) in making the final decision [33], [32]. The LR algorithm also plays an important role in the process of predicting diabetes, as it identifies the independent variables and classifies them on the x- and y-lines, and then measures the probabilities of an event's [34],[31]. One of the recently discovered options for machine learning is the XGBoost model which also counts on ensemble learning methodology; it can deal with unbalanced datasets classes by measuring the loss function and resolve the problem of overlearning using grading and voting [27], [28], [29].

A medical diagnosis of diabetes mellitus is one of the challenges in the medical field. Patients' information may include age, body mass, triceps skinfold thickness, serum insulin, plasma glucose concentration, diastolic blood pressure, and other factors. Based on these elements, the decision will

made. The decision-making process is drawn out and takes weeks or months, making the doctor's job incredibly hard, but with the help of new technologies, it will be easier; consequently, machine learning techniques are a crucial solution [9].

Today, an extensive selection of medical datasets that are helpful for research in fields of medical science are easily accessible [10]. According to all this information and background about diabetes mellitus disease and the most prominent techniques used to predict it, we propose a novel stacking ensemble model for the prediction of DM utilizing a combination of machine learning models.

The Contributions of this paper are as follows:

- Developing a stacking ensemble model for predicting Diabetes Mellitus using a combination of machine learning models.
- Merging the RF and LR models as base learners and the XGB model as a meta-learner in building the proposed stacking model.

This paper is arranged as follows: Section II provides the Related work; Section III covers the material and methodologies of our study; Section IV illustrates the experimental setup for our proposed model; and Section V the performance measures. Our results and discussion are discussed in Section VI and VII respectively. Finally, the Conclusion has been addressed in Section VIII.

II. Related Work

Ensemble learning is a computational and statistical approach. Mimicking how people learn social skills by experimenting with different viewpoints before making the final decision. A Set of machine learning models combines choices and provides more robust and accurate predictions [11-12].

Gollapalli et al. [13] proposed a novel stacking ensemble model using machine learning to detect three-types of diabetes mellitus: T1DM, T2DM, and Pre-diabetes. Empirical results showed that the proposed model could predict with 94.48 percent accuracy, 94.48 percent recall, 94.70 percent precision, and 0.917 percent Cohen's kappa score. After observation, the most critical features of predicting T1DM, T2DM are: Sex; human gender A1c: measures the amount of sugar bonded to the hemoglobin protein in the blood; TG: The blood triglyceride level of the patient; LDL: Low-density lipoprotein, or LDL, is a measure of the quantity of harmful cholesterol; AntiDiab: A blood sugar-lowering oral medicine used to combat diabetes; Albumin: The amount of protein; Insulin, Injectable, Nutrition, Education. However, the study needed more ML classifiers and deep learning models to increase prediction accuracy.

Dutta et al. [14] emphasize that using an ML-based ensemble model in predicting DM is critical in ensuring more accurate predictions. Also, exploring deep learning techniques and applying them with an ensemble learning approach is recommended. Stacking is an ensemble method that employs a meta-model in which a novel classifier integrates multiple weak learners to predict the target variable [13].

Ganie and Malik [15] discussed the various ensemble learning methods, such as the Bagging method, in predicting T2DM based on lifestyle indicators. The synthetic minority oversampling technique is used for dataset class balancing. Furthermore, the results are validated using the Cross-Validation technique. Researchers and practitioners use the cross-validation technique for the model-building process to remove biases.

Laila et al. [16] studied efficient ensemble algorithms for predicting diabetic risks in the early stages, using Seventeen features gathered from the UCI of various datasets. This research used predictive models like (AdaBoost, Bagging, and Random Forest) were utilized to evaluate accuracy, recollection, and F1-score. Overall, the RF ensemble methodology had the highest accuracy (97 %), whereas AdaBoost and Bagging had lesser accuracy.

Javale and Desai [17] concentrated on an ensemble technique for healthcare information analytics employing machine learning through unbalanced dataset approaches, synthetic minority over-sampling, plus adaptive synthetic over-sampling. Using other analysis techniques such as the train-test, the K-folds, and the repeat train-test. The average Stacking-C technique was used to execute an ensemble strategy on the diabetes dataset, which included K-Nearest Neighbors (KNN), Support vector machines (SVM), RF, Naïve Bayes (NB), and logistic regression classifiers. The Synthetic Minority Oversampling Technique (SMOTE) reduces False Negative counts with more precision. An ensemble method facilitates appropriate decision-making by providing a more profound knowledge of the implementation. Rather than just comparing the classifiers' outputs produced for various performance measures, choosing the optimal ensemble technique for the application is always preferable. The fundamental challenge in healthcare information analytics is unbalanced datasets, which might be a critical factor for an ensemble technique in healthcare data analytics.

Singh et al. [18] suggested an ensemble-based approach for diabetes prediction called eDiaPredictTo forecast diabetes status in patients, it employs ensemble modeling, which consists of an ensemble of multiple machine learning algorithms such as XGBoost, RF, SVM, NN, and DT. The minimalizing error value and lowest weighted coefficient of eDiaPredict have all been tested. The suggested approach's usefulness is shown using the PIMA Indian medical dataset, which has an accuracy of 95% percent. The stacking ensemble combines the predictions of many ML models to get the maximum accuracy achievable compared to the conventional models. It leverages a single model known as a meta-model to diagnose the optimum mix of expectations from the basic models. The stacking ensemble contains two stages, level 0 and level 1. The former employs heterogeneous ML models known as base learners. In contrast, the latter uses a single model known as a meta learner, whose purpose is to unify the predictions of the basis learners. To predict T2DM and alert patients in advance to decrease the risk factor and intensity associated with diabetic diseases.

Geetha and Prasad [19] suggested a hybrid ensemble model. For the decision tree, they employed ensemble approaches such as bagging with "random forest" and Adaboost and supervised classification algorithms like Naive Bayes. Merge different two or more models improves performance by increasing the accuracy and precision of predictions. Joshi et al. [31] focused on predicting Type 2 diabetes in Pima Indian women using a logistic regression (LR) model and a decision tree, and the accuracy of the proposed model was 78% percent.

Patil et al. [20] proposed a framework for T2DM prediction that uses a stacking-based ensemble with a "non-dominated sorting genetic algorithm" method. The main objective is to reduce the time elapsed between diabetes diagnosis and medical evaluation. The suggested NSGA-II stacking method was compared to Boosting, Bagging, RF, and Random Sub-space approaches. The stacking ensemble methodology has outperformed all other traditional ensemble approaches. Findings indicate that the NSGA-II stacking approach performs better over other conventional ensemble methods with an accuracy of 81 percent.

Syed and Khan [21] created an ML-Based System for Predicting the Risk of (T2DM), which is a web-based prediction model that uses Azure ML to estimate the risks of Type 2 diabetes. The results show that the suggested model can accurately predict the risks of Type 2 diabetes by 82 percent. The geographical range of this study was restricted since it primarily focused on the western portion of Saudi Arabia for the validation procedure. Table I explains the most important studies according to limitations and Advantages, Data Sources.

III. Material and Methodologies

This research proposes a stacking ensemble model to predict diabetes mellitus. The proposed model relies on two essential levels of construction. The first level is called (base learners). At this level, a combination of machine learning models is prepared, trained, and produced predictions that are entered as inputs to a new model/classifier that learns from these inputs to make the final prediction (the meta-learner), the second level. We have selected the logistic regression model, Random Forest, as base learners with their distinction ability in binary classification processes. We select The XGBoost model as a meta-learner, which contributes positively to dealing with imbalanced dataset classes by minimizing the loss function and increasing the weight of the classified incorrect classes. The optimization GridSearchCV technology is applied to get the best possible results by the base learners and the meta learner; it uses a grid search of hyperparameters tuning for each model and extracts the best results. In our proposed model, we have included the cross-validation technique with default five-fold iterations to get optimal results. Also, we applied this technique on each of the base learners: Random Forest, Logistic regression, via the Optimizer GridSearchCV to get the best results by using the sci-kit-learn library, which provides a random split into training and test sets can easily calculate with the train_test_split assistant function. Each model starts with using K-1 of the folds. Fig. 1 describes the methodology for our proposed stacking model:

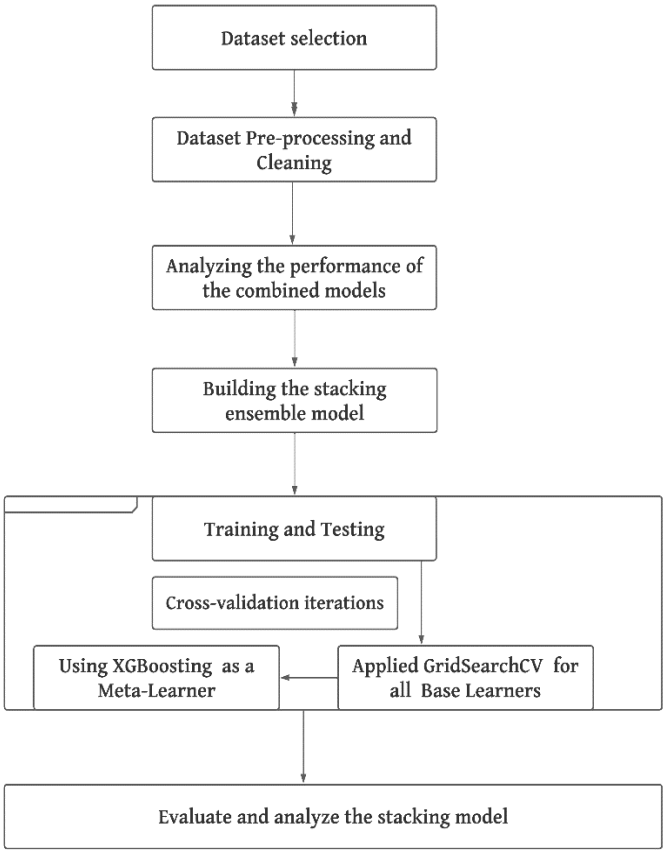

Fig. 1. Proposed stacking model.

A. Stacking

Stacking is an ensemble learning method that uses a meta-model where a new classifier integrates several individual-based learning predictions to get the best combined predictions. The stacking method has two levels in the building; level 0 (base-learners) combine heterogeneous models that are fitted and trained on a dataset, then the results will be fed as input to the meta-learner at the next level. The level 1 (meta learner) learns how to combine the predictions from the base models and provide robust and high-accuracy predictions [13]. We built our proposed model based on this stacking ensemble method with more of our contributions, like utilizing of cross-validation technique for all learners and leveraging the GridSearchCV hyperparameters tuning technique for the base learners. Fig. 2 illustrates the ensemble stacking methodology. Where m*n means that n of number k-folds Cross-validations of training dataset that will go cross all base learners' models, and m*M means that m of numbers of predictions coming from a number of M base learners will send to the next meta-learner model as inputs and then he learns from all of these predictions how to predict the final prediction.

TABLE I. IMPORTANT STUDIES IN PREDICTING DIABETES MELLITUS

Ref.	Algorithms	Data Sources	Advantages	Limitations
M. Gollapalli et al, 2022	Support Vector Machine, K-nearest Neighbor and Decision Tree.	Hospital (KFUH). Saudi Arabia	Use of Cross-validation technique in training the models, which leads to increased performance in prediction.	Need for using more and different machine learning models to improve results.
A. Dutta et al, 2022	Decision tree, Random Forest, Extreme Gradient Boosting, Light gradient boosting machine.	DDC dataset Bangladesh.	Use of Hyperparameter Optimization (Grid Search) for tuning the models.	Need for a large dataset to improve results.
A. Singh et al, 2021	Extreme Gradient. Boosting, Random. Forest, Support Vector Machine, Neural Network, and Decision tree.	PIMA Indian diabetes	Use of Recursive Feature Elimination (RFE) for feature space reduction in the dataset	Application of the proposed model in medical life tests.
A. H. Syed and T. Khan, 2020	Decision forest.	PIMA Indian diabetes.	Use of SMOTE technique for balancing dataset classes. which leads to avoiding overfitting.	Geographical scope of the study.
S. M. Ganie and M. B. Malik, 2022	Bagged Decision Trees, Random Forest, Extra Trees, AdaBoost, Stochastic Gradient Boosting, Voting.	Manually	Using the seaborn-Facet Grid method to visualize the dataset elements.	Develop an application for the proposed model to predict type 2 diabetes.
S. Härner and D. Ekman (2022)	Decision tree, Naïve Bayes.	PIMA Indian diabetes.	Comparing ensemble methods for predicting diabetes mellitus.	Need for using hyperparameters search optimizer to improve model results.

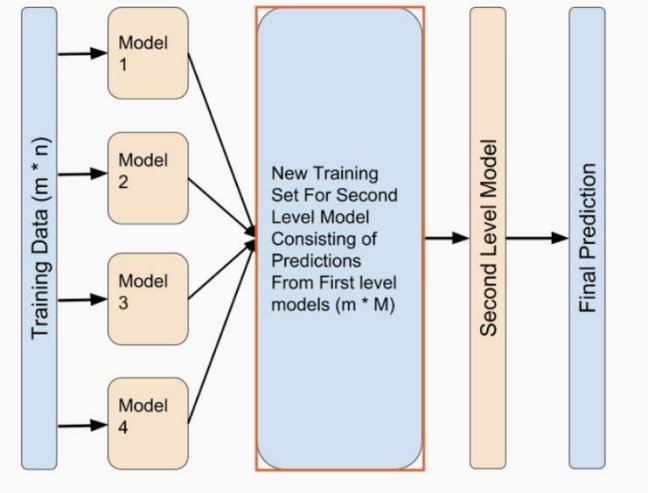

Fig. 2. Stacking ensemble method.

B. GridSearchCV

GridSearchCV is a widely used technique in machine learning and deep learning model development. It helps select the best hyperparameter values for a specific model. Hyperparameters determine the model's behavior and configuration, such as the number of layers and batch size in neural networks or the depth of trees in decision trees.

GridSearchCV works by systematically testing different combinations of hyperparameters and evaluating their performance through cross-validation. This involves defining a set of possible values for each hyperparameter, training and testing the model with each combination, and ultimately selecting the combination that yields the best performance [14]. This process enhances the model's performance and prevents overfitting by evaluating it on separate training and testing data sets [14]. This approach significantly enhances model performance and mitigates overfitting by utilizing cross-validation, assessing the model on distinct training and testing datasets. [14].

C. Logistic Regression

Logistic regression is a machine learning algorithm used for binary classification. It is specifically designed to build models that can predict specific classifications. Despite its name, it is not used for predicting logistic events, but rather for classification based on the logistic (sigmoid) function [23].

Logistic regression solves the binary classification problem by classifying examples into one of two classes (e.g., class 0 or class 1) based on a set of independent variables (features). The algorithm uses the logistic function to model and determine the probability of an example belonging to the positive class (class 1) [24], [25]. In logistic regression, the results of the logistic function are transformed into probabilities using the sigmoid function. This conversion ensures that the probabilities range between 0 and 1. These probabilities represent the estimated classification probability, which is used to make the final classification decision [31]. Logistic regression is widely used across various fields, including data science, business analytics, medicine, and text classification. It is valued for its simplicity and its ability to handle large datasets efficiently [23].

D. Random Forest

Random Forest is a machine learning technique used for classification and prediction. It is a significant algorithm in optimization and diversification for classification and prediction models [32].

Random Forest is based on a collection of Decision Trees, a decision Tree divides data into categories by making sequential choices [32]. Each choice splits the data into subsets based on specific questions about the available variables. Random Forest creates a set of Decision Trees randomly by:

1) Randomly selecting samples with replacement from the original dataset (training data) for each decision tree.

2) Randomly selecting a subset of variables to build each tree.

Once the set of Decision Trees is constructed, Random Forest combines the individual tree predictions through voting or averaging to make a final decision for classification or prediction. Random Forest offers advantages such as model diversity and reducing overfitting, which occurs when the model becomes overly specialized to the training data. It also uses variable importance information to assess the impact of each variable on classification, providing valuable insights into the data [32]. Random Forest is widely used in various applications including image classification, word recognition, price prediction, and environmental analysis [32].

E. Extreme Gradient-Boosting

The gradient-boosting decision tree (GBDT) is the foundation of XGBoost, which was proposed by Tianqi Chen et al. [26]. A gradient-boosting algorithm built on a decision tree is called GBDT. Gradient boosting is an ensemble learning method that combines several weak classifiers into a stronger classifier during training. The objective of computing negative gradients is to enhance the following training cycle by minimizing the loss function and increasing the weight of the classified incorrectly classes. In contrast to GBDT, XGBoost incorporates a regularization technique to minimize model complexity, improve loss function smoothness, and prevent overfitting. To improve gradient boosting, locate the best-split solution, and promote scalability and efficiency, an approximation approach is also applied. XGBoost additionally enables parallel operations and an early stop to speed up the model operation. The model tree can stop to speed up training when the forecast result reaches the optimum. The model's classification accuracy can also be increased with XGBoost [27]. Zhao et al. [28] stated that XGBoost could effectively prevent the training model from over-fitting. Secondly, embedded parallel processing allows a faster learning speed.

Moreover, the XGBoost classifier can learn from imbalanced training data by setting class weight and taking ROC as evaluation criteria. XGBoost is one of the best classifiers for dealing with imbalanced datasets when the dataset classes with less variance [29]. Consequently, in this research, we chose it as the meta-learner of our proposed stacking model. The Extreme Gradient Boosting (XGBoost) is a machine learning algorithm used for classification and prediction tasks. It is an evolution of gradient boosting, combining multiple simple models into a strong and accurate model to enhance performance and accuracy.

XGBoost creates a sequence of weak models, like shallow decision trees, and boosts their performance. This boosting process focuses on the data predicted incorrectly by previous models, improving the overall performance of the model. Key features of XGBoost include:

1) Performance enhancement: XGBoost is known for achieving superior performance in various classification and prediction domains.

2) Multi-objective versatility: It can be used for both classification problems and numerical value prediction.

3) Overfitting prevention: Boosting parameters can be adjusted to limit overfitting and prevent excessive learning from training data.

4) Time and resource optimization: XGBoost strikes a balance between speed and accuracy, optimizing performance and resource utilization.

5) Handling missing data: XGBoost intelligently handles missing values without extensive preprocessing.

Overall, XGBoost is a powerful and popular tool in machine learning. It can effectively solve complex problems and improve predictive model performance.

F. Cross-Validation

Cross-validation is a popular data resampling method for evaluating a predictive model's generalization capacity and preventing overfitting by splitting the dataset into k-folds during training and testing iterations. The term "fold" here describes the quantity of generated subsets. The learning set's cases are randomly sampled for this division without being replaced. k-1 subsets comprise the training set and are used to train the model. The quality of this technology lies in storing unseen data at each new n-fold, which makes the prediction result more accurate. The model's performance is evaluated after being applied to the final subset, the "unseen dataset." This process is repeated until each of the k subsets has acted as a validation set [22]. Fig. 3 illustrates the cross-validation technique [30]. To achieve the best results, we used the cross-validation technique with the default four-fold iterations in our stacking model. We also used this technique on each of the base learners: RF, LR, and the Meta-Learner XGBoost via the Optimizer GridSearchCV to achieve the best results by using the sci-kit-learn library, which provides a random split into training and test sets that can be easily calculated with the train_test_split assistant function.

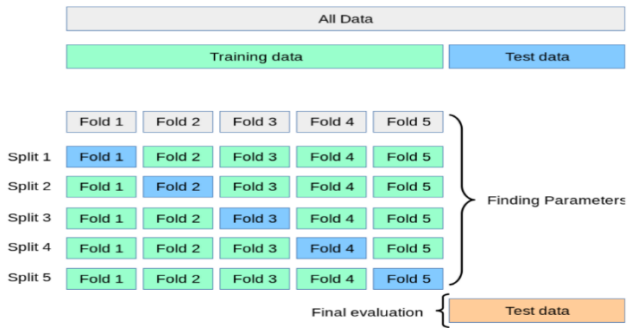

Fig. 3. Cross-Validation technique.

G. Study Dataset

The PIMA dataset, also known as the PIMA Indian Diabetes dataset, is a well-known dataset used in machine learning and data mining. It is named after the Pima Native American tribe in Arizona, USA. This dataset is commonly used for classifying the onset of diabetes in individuals by using medical diagnostic measurements [13]. It's available on Kaggle worldwide datasets repository, link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

This dataset consisted of 268 diabetic (positive = 1) and 500 non-diabetic (negative = 0) patients with eight Features presented below and in Table II. [13]:

1) Pregnancies: Number of pregnancies.

2) Glucose: Plasma glucose concentration measured two hours after an oral glucose tolerance test.

3) Blood Pressure: Diastolic blood pressure (mm Hg).

4) Skin Thickness: Triceps skinfold thickness (mm).

5) Insulin: Serum insulin level measured two hours after consumption (mu U/ml).

6) BMI: Body mass index (weight in kg / (height in meters) ^2).

7) Diabetes Pedigree Function: A function that estimates the likelihood of diabetes based on family history.

8) Age: Age in years.

TABLE II. PIMA DATASET INFORMATION

Data columns (total 9 columns):			
#	Column	Non-Null Count	Datatype
0	Pregnancies	768 non-null	int64
1	Glucose	768 non-null	int64
2	Blood Pressure	768 non-null	int64
3	Skin Thickness	768 non-null	int64
4	Insulin	768 non-null	int64
5	BMI	768 non-null	float64
6	Diabetes Pedigree Function	768 non-null	float64
7	Age	768 non-null	int64
8	Outcome	768 non-null	int64

Pima dataset is commonly used to demonstrate various machine learning techniques, such as logistic regression, decision trees, support vector machines, and neural networks, for predicting the likelihood of diabetes based on these medical measurements. It's important to note that while the PIMA dataset is valuable for educational purposes and experimenting with machine learning algorithms, it is relatively small and has limitations, such as missing values and potential biases. Therefore, caution should be exercised when drawing conclusions or developing predictive models solely based on this dataset in real-world applications.

Statistical analysis provides essential tools for visualizing and understanding the dataset pattern to improve the data pre-processing and modeling process. Fig. 5 presents the statistical information of the features and data types found in the PIMA dataset. Table III shows the distribution of the features based on count, mean, standard deviation, maximum value, minimum value, and the percentile/quartile of each feature. The correlation coefficient has been used to measure the feature relationships in Fig. 6, and finally, outcomes values are presented in Fig. 4.

IV. EXPERIMENTAL SETUP

In this research, we used Jupyter Notebook to build the stacking ensemble model, using Microsoft Intel(R) Core i5-1035G7 CPU 1.20GHz and 8 Giga RAM. The public Pima Dataset has been selected, pre-processed, and cleaned up from a few defects, such as zero values in features columns, using the arithmetic mean for each column. The base learners' models were initialized with a Random Forest model using the Grid-search Hyperparamets Tunner through the following Hyperparameters: 'bootstrap training,' 'max of samples training,' 'max_features,' 'min_samples_leaf,' 'min_samples_split,' 'n_estimators.' Moreover. The second base learner, Logistic regression: "C," np. Logspace, "penalty":12. To address the problem of an imbalanced data set, which causes overfitting and inconsistent results, we applied the Extreme Gradient Boosting model as a meta-learner, which is counted on an ensemble learning method that allows us to deal with unbalanced dataset classes. A cross-validation technique was implemented for the proposed stacking model using default 5-k folds; they were also included through the GridSearchCV of hyperparameters for the base learners and the meta learner. Finally, the proposed stacking model was verified on a new dataset containing 100,000 records.

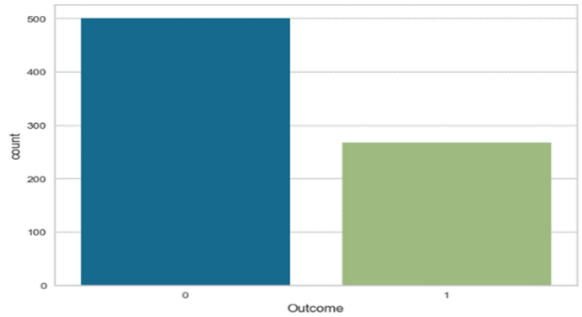

Fig. 4. Pima dataset outcome targets.

TABLE III. STATISTICAL DISTRIBUTION OF PIMA DATASET FEATURES

	Pregnancies	**Glucose**	**Blood Pressure**	**Skin Thickness**	**Insulin**	**BMI**	**Diabetes Pedigree Function**	**Age**	**Outcome**
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

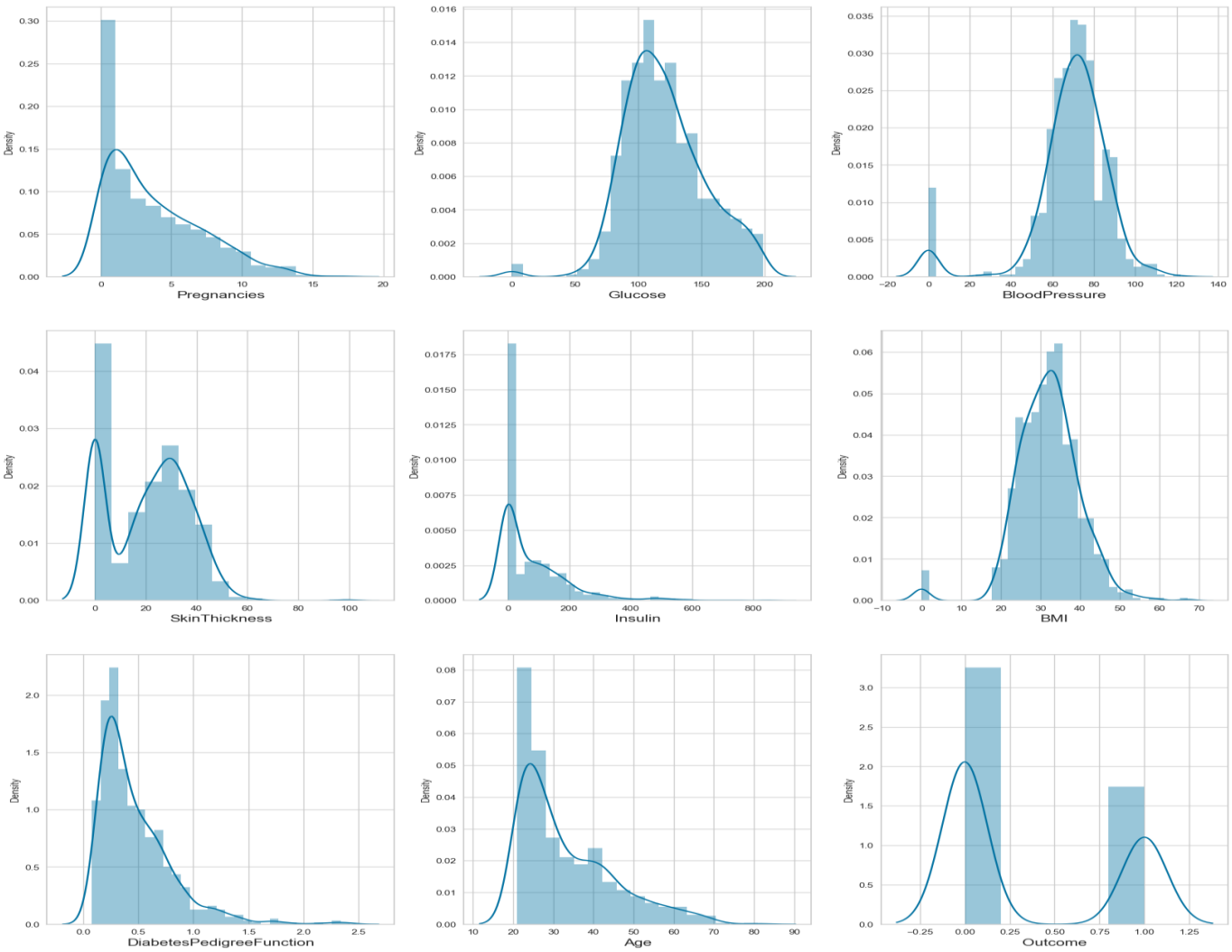

Fig. 5. Pima Dataset features chart .

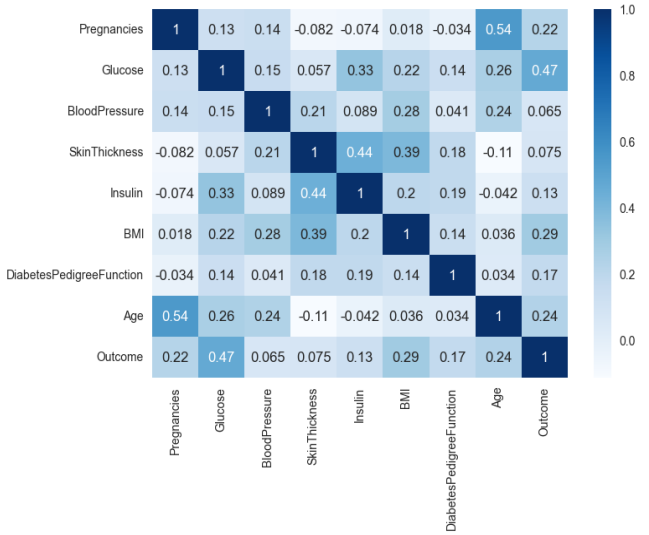

Fig. 6. Pima dataset features correlation heatmap.

V. Performance Measure

To assess the performance of classification models in machine learning and data analysis, we utilize the following metrics:

Accuracy: This metric represents the ratio of correctly predicted samples to the total number of samples. It measures the model's ability to accurately classify both positive and negative cases. However, it's important to note that accuracy can be misleading when dealing with imbalanced classes, as high accuracy can be achieved without focusing on positive classification [13].

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

The output is either diabetic (+dm) or not diabetic (-dm).

- True positive (TP): Prediction is +dm and X is diabetic.
- True negative (TN): Prediction is -dm and X is not diabetic.

- False positive (FP): Prediction is +dm and X is not diabetic.
- False negative (FN): Prediction is -dm and X is diabetic.

Precision: Precision measures the accuracy of predicting positive cases. A high precision value indicates that the model correctly classifies cases as positive when it claims they are [13].

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

The output is either diabetic (+dm) or not diabetic (-dm)

- True positive (TP): Prediction is +dm and X is diabetic.
- False positive (FP): Prediction is +dm and X is not diabetic.

Recall: Also known as sensitivity or true positive rate, recall measures the model's ability to identify all available positive cases. A high recall value signifies that the model can identify most positive cases [13].

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The output is either diabetic (+dm) or not diabetic (-dm)

- True positive (TP): Prediction is +dm and X is diabetic.
- False negative (FN): Prediction is -dm and X is diabetic.

Cohen's Kappa Score: This metric measures the agreement between two raters. In the context of evaluating classification models, Cohen's Kappa score gauges the agreement between the model's classification and the actual classification. It proves particularly useful when dealing with imbalanced classes or when the model randomly selects between classes [13].

$$CKS = \frac{P_0 - P_e}{1 - P_e} \quad (4)$$

TABLE IV. THE STACKING MODEL RESULTS

	Model	Score
0	Random Forest	0.750730
1	Logistic Regression	0.773706
2	Stacking Model	0.828571

TABLE V. BASE AND META LEARNERS RESULTS

Targets	Precision	Recall	F1-score support	support
0	0.80	0.95	0. 87	41
1	0.90	0. 66	0. 76	29
Accuracy			0.83	70
Macro avg	0.85	0.80	0.81	70
Weighted avg	0.84	0.83	0.82	70

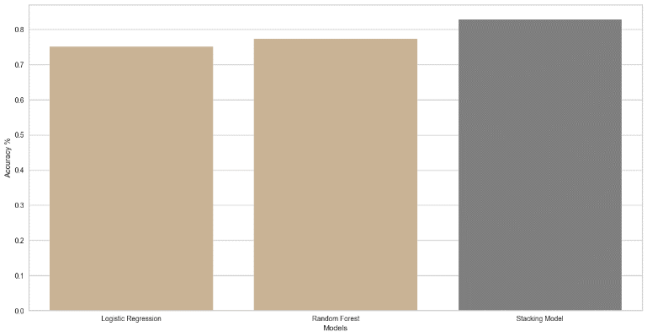

Fig. 7. Base and meta learners results.

where, P_0 represents the accuracy of the models and P_e denotes the agreement between the predicted and actual labels [13].

These metrics aid in comprehending the performance of classification models and identifying their strengths and weaknesses. It is recommended to employ a variety of these metrics to obtain a comprehensive understanding of the model's performance.

A. Validation Dataset

We validated our proposed stacking model performance on a new (binary classification outcomes) diabetes dataset. Diabetes prediction dataset (DPD) is a public dataset consisting of electronic health records (EHRs) that contain digital copies of health records for patients' medical history, diagnosis, therapy, and outcomes. EHRs data is gathered and kept by healthcare providers such as medical centers and hospitals as part of their usual clinical practice. DPD has approximately 100,000 patient records, contributing to significantly measuring the proposed model performance. Its available on Kaggle worldwide datasets repository, link: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset. DPD dataset features are shown in Table VI; moreover, the correlation heatmap between features is displayed in Fig. 8, whereas Table VII shows the results.

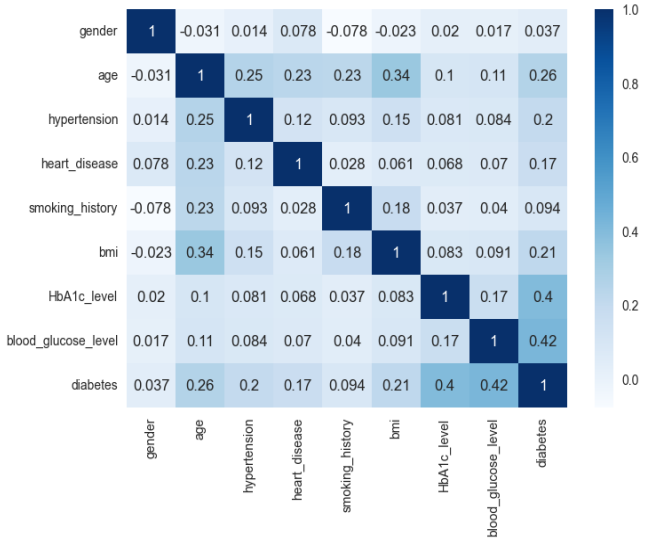

Fig. 8. DPD Dataset features correlation heatmap.

TABLE VI. DPD DATASET INFORMATION

Data columns (total 9 columns):		
Column	Non-Null Count	Datatype
Gender	100,000 non-null	object
Age	100,000 non-null	float64
hypertension	100,000 non-null	int64
Heart disease	100,000 non-null	int64
Smoking history	100,000 non-null	object
BMI	100,000 non-null	float64
HbA1c_level	100,000 non-null	float64
Blood_glucose_level	100,000 non-null	int64
Outcomes	768 non-null	int64

TABLE VII. THE STACKING MODEL RESULTS ON DPD DATASET

	Model	Score
0	Random Forest	0.95064
1	Logistic Regression	0.97012
2	Stacking Model	0.97160

VI. RESULTS

In this research, we built a stacking ensemble model to predict diabetes mellitus using a combination of machine learning models; where random forest, logistic regression models were applied as base learners and the extreme gradient Boosting model as meta learner, techniques such as cross validation and GridSearchCV were applied. We also replaced the zeros in the Pima dataset with values (median - mean) according to the types of data distribution with features columns (normal - skewed), as mentioned in [71] that if we remove zero values, the performance will improve. We obtained an accuracy of 83% in predicting diabetes mellitus with Pima dataset, and we also verified the efficiency of the proposed model on a large dataset containing approximately 100,000 records, with accuracy of 97%, where kapa Cohen score was 61% on Pima dataset, and 78% on DPD dataset. More details are discussed in the next paragraph. We observe that our proposed ensemble stacking model for predicting diabetes covers the shortcomings mentioned in Table I, such as the study of S. Härner and D. Ekman (2022) regarding the need for using hyperparameters search optimizer to improve model results. Moreover, a study of H. Syed and T. Khan (2020) about the geographical scope of the study dataset, where we used two different dataset scopes. The Table IV shows the detailed results of the proposed stacking model. In addition, Table V, Fig. 7 shows the base and meta learners results.

VII. DISCUSSION

A. Results with XGBoost and GridSearchCV

In this experiment, we utilized a combination of ML and DL classifiers to predict diabetes mellitus. Fig. 1 illustrates the stacking model methodology in this experiment, where we initiated each of the RF, LR models as base learners and the XGB classifier as a Meta-learner. At the same time, we used GridSearchCV Hyperparameters optimizer to find the optimal results for the Random Forest classifier using the following hyperparameters: bootstrap, max_features, min_samples_leaf, n_samples, split,n_estimators Moreover. The second base learner, Logistic regression: "C," np. Logspace, "penalty":12. To address the problem of an imbalanced data set, which causes overfitting and inconsistent results, we applied the Extreme Gradient Boosting model as a meta-learner, which is counted on an ensemble learning method that allows us to deal with unbalanced dataset classes. A cross-validation technique was implemented for the proposed stacking model using default 5-k folds; they were also included through the GridSearchCV of hyperparameters for the base learners and the meta learner. The results were as follows: The prediction accuracy of the stacking ensemble model on Pima dataset is 83%, kapa Cohen score 61%, where on DPD dataset was 97% accuracy and 78% kapa Cohen score. Table IV shows the results in detail. Fig. 9 displays the differences in results between Pima and DPD datasets.

B. Comparative Analysis with Existing Work

1) First study: S. Härner and D. Ekman (2022) [34] proposed an ensemble stacking model for predicting diabetes using a combination of machine learning models, including (Decision Tree and Naive Bayes models). The Pima dataset was used in this study, and the results indicated that the proposed stacking model can predict diabetes with 75.56% accuracy. In addition, it was mentioned that there were limitations during the study, such as not using an optimizer to search in hyperparameters to find the best results for base learners in the stacking model.

2) Second study: Patil et al. (2023) [20] Suggested an ensemble stacking model for predicting diabetes, using a combination of machine learning models such as (decision tree, naïve Bayes (NB), multilayer perceptron (MLP), SVM, and KNN). The Pima dataset was also used in this study. The results indicated that the stacking model can predict diabetes with 81.9% accuracy. In addition, we noticed that they never mentioned the cross-validation technique during the proposed methodology, which plays an essential role in building the stacking model. Moreover, no optimizer was used in searching the hyperparameters while training the base learners' models to get better results.

3) Third study: Lei Qin (2022) [35] devised an ensemble stacking approach to predict diabetes. They amalgamated various machine learning models—Logistic Regression, K-Nearest Neighbors, Decision Trees, Gaussian Naive Bayes, and Support Vector Machine (SVM). Utilizing the Pima dataset, their findings revealed that the stacking model achieved an 81.63% accuracy in diabetes prediction. However, the absence of an optimizer for hyperparameter tuning during base learner model training might have hindered the quest for better outcomes. Additionally, the limited size of the dataset posed a challenge, potentially impacting the attainment of optimal results.

4) Forth study: Kumari et al. (2021) [36] suggested an ensemble soft voting model for predicting diabetes, using a combination of machine learning models such as (Random

Forest (RF), logistic regression (LR), and Naive Bayes (NB)). The Pima dataset was also used in this study. The results indicated that the soft voting model can predict diabetes with 79.04% accuracy. Furthermore, it's important to highlight that the proposed methodology overlooked the inclusion of cross-validation, a crucial technique integral to ensuring robustness by assessing the performance of individual models across various subsets of the data, thereby refining their predictions' collective contribution to the ensemble. Additionally, the absence of an optimizer in the pursuit of hyperparameter tuning during the training of base learner models might have impacted the potential for achieving superior results.

We observe that our proposed ensemble stacking model outperforms in predicting diabetes accuracy compared to other proposed models in the [20], [34], [35], [36] studies, in our approach, we leveraged the GridsearchCV optimizer to search for the best hyperparameters for our base learners. Interestingly, this optimization technique wasn't utilized in either the First Study or the Second Study. This optimization significantly boosted our base learners' learning process, leading to extracting the most optimal results possible. Furthermore, the second and fourth studies overlooked the utilization of cross-validation—a critical technique for evaluating a predictive model's generalization capacity. In contrast, our model applied this method, dividing the dataset into k-folds during both training and testing. This implementation effectively evaluated and prevented overfitting, significantly enhancing our prediction model's performance.

Table VIII meticulously delineates and highlights the disparities and advantages between these studies, emphasizing the significant enhancements our approach brings to the table in comparison to the methodologies adopted in the First Study and the Second Study.

TABLE VIII. COMPARISON WITH EXISTING STUDIES

Authors	Techniques used	Dataset	Accuracy
S. Härner and D. Ekman (2022)	stacking ensemble approach (Decision tree (DT), Naïve Bayes (NB)), Cross-validation	Pima dataset	75.56%
Patil et al (2023)	Stacking ensemble approach (Decision tree (DT), Naïve Bayes (NB), multilayer perceptron (MLP), SVM, and KNN)	Pima dataset	81.9%
Lei Qin (2022)	Stacking ensemble approach (Logistic Regression, K-Nearest Neighbors, Decision Trees, Gaussian Naive Bayes, and Support Vector Machine (SVM))	Pima dataset	81.63%
Kumari et al (2021)	Soft voting ensemble approach (Random Forest (RF), Logistic regression (LR), and Naive Bayes (NB))	Pima dataset	79.04%
Our proposed model	Stacking ensemble approach (Random Forest, Logistic Regression, XGboost) GridSearchCV, Cross-validation.	Pima Dataset	83%
Our proposed model on the validation dataset	Stacking ensemble approach (Random Forest, Logistic Regression, XGboost) GridSearchCV, Cross-validation.	DPD Dataset	97%

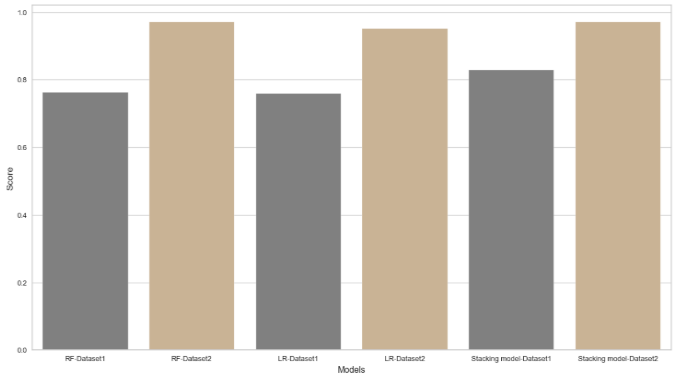

Fig. 9. Comparison results between pima and DPD datasets.

VIII. CONCLUSION

Diabetes mellitus (DM) is a common disease that threatens the health of society, causing many serious diseases such as kidney failure, heart disease, and blindness. In this research, we proposed a novel stacking ensemble model to predict diabetes mellitus using a Pima dataset and combined machine learning models, where we used the Random Forest (RF) and Logistic Regression (LR) as base learners models and XGBoost as a Meta-Learner model. Moreover, we applied the cross-validation technique to get the optimal results in the RF, LR, models through the Grid Search optimizer technique. To avoid the problem of an imbalanced dataset, which causes overfitting and inconsistent results, we applied the XGBoost model as a meta-learner. However, the dataset has been cleaned from zero values that harm the prediction result, which was illogical to have zero values on some columns, like glucose in the blood. To address this problem, we replaced zero values with median and mean values based on the type of distribution (normal - skewed). The results indicate that our proposed stacking model can predict diabetes mellitus with an accuracy of 83% with the Pima dataset, and 97% on the DPD dataset. As recommendations, our stacking model can be applied in a diagnostic application for diabetes mellitus; in addition, it can be tested on a new huge and diverse dataset to obtain more accurate results. Moreover, we can use deep-learning models to generate new patterns that help us diagnose DM robustly, which also can happen with different types of diabetes, such as type 1 and type 2 diabetes and gestational diabetes.

REFERENCES

[1] H. Sone, 'Diabetes Mellitus', in Encyclopedia of Cardiovascular Research and Medicine, R. S. Vasan and D. B. Sawyer, Eds., Oxford: Elsevier, 2018, pp. 9–16. doi: https://doi.org/10.1016/B978-0-12-809657-4.99593-0.

[2] T. Andoh, 'Subchapter 19A - Insulin', in Handbook of Hormones, Y. Takei, H. Ando, and K. Tsutsui, Eds., San Diego: Academic Press, 2016, pp. 157-e19A-3. doi: https://doi.org/10.1016/B978-0-12-801028-0.00148-3.

[3] J. Hippisley-Cox and C. Coupland, 'Diabetes treatments and risk of amputation, blindness, severe kidney failure, hyperglycaemia, and hypoglycaemia: open cohort study in primary care', BMJ, p. i1450, Mar. 2016, doi: 10.1136/bmj.i1450.

[4] A. N. Baanders and M. J. W. M. Heijmans, 'The Impact of Chronic Diseases: The Partner's Perspective', Family & Community Health, vol. 30, no. 4, pp. 305–317, Oct. 2007, doi: 10.1097/01.FCH.0000290543.48576.cf.

[5] P. Saeedi et al., '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, vol. 157, p. 107843, Nov. 2019, doi: 10.1016/j.diabres.2019.107843.

[6] C. V. A. Collares et al., 'Transcriptome meta-analysis of peripheral lymphomononuclear cells indicates that gestational diabetes is closer to type 1 diabetes than to type 2 diabetes mellitus', Mol Biol Rep, vol. 40, no. 9, pp. 5351–5358, Sep. 2013, doi: 10.1007/s11033-013-2635-y.

[7] A. E. Butler and D. Misselbrook, 'Distinguishing between type 1 and type 2 diabetes', BMJ, p. m2998, Aug. 2020, doi: 10.1136/bmj.m2998.

[8] World Health Organization, Global report on diabetes. Geneva: World Health Organization, 2016. Accessed: Jan. 25, 2023. [Online]. Available: https://apps.who.int/iris/handle/10665/204871

[9] A. Thammano and A. Meengen, 'A New Evolutionary Neural Network Classifier', in Advances in Knowledge Discovery and Data Mining, T. B. Ho, D. Cheung, and H. Liu, Eds., in Lecture Notes in Computer Science, vol. 3518. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 249–255. doi: 10.1007/11430919_31.

[10] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[11] T. G. Dietterich, 'Ensemble Methods in Machine Learning', in Multiple Classifier Systems, in Lecture Notes in Computer Science, vol. 1857. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1.

[12] L. I. Kuncheva, Combining pattern classifiers: methods and algorithms, Second edition. Hoboken, NJ: Wiley, 2014.

[13] M. Gollapalli et al., 'A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM', Computers in Biology and Medicine, vol. 147, p. 105757, 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.105757.

[14] A. Dutta et al., 'Early Prediction of Diabetes Using an Ensemble of Machine Learning Models', IJERPH, vol. 19, no. 19, p. 12378, Sep. 2022, doi: 10.3390/ijerph191912378.

[15] S. M. Ganie and M. B. Malik, 'An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators', Healthcare Analytics, vol. 2, p. 100092, Nov. 2022, doi: 10.1016/j.health.2022.100092.

[16] U. e Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, 'An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study', Sensors, vol. 22, no. 14, p. 5247, Jul. 2022, doi: 10.3390/s22145247.

[17] D. Pankaj Javale and S. Suhas Desai, 'Machine learning ensemble approach for healthcare data analytics', IJEECS, vol. 28, no. 2, p. 926, Nov. 2022, doi: 10.11591/ijeecs.v28.i2.pp926-933.

[18] A. Singh, A. Dhillon, N. Kumar, M. S. Hossain, G. Muhammad, and M. Kumar, 'eDiaPredict: An Ensemble-based Framework for Diabetes Prediction', ACM Trans. Multimedia Comput. Commun. Appl., vol. 17, no. 2s, pp. 1–26, Jun. 2021, doi: 10.1145/3415155.

[19] G. Geetha and K. M. Prasad, 'An Hybrid Ensemble Machine Learning Approach to Predict Type 2 Diabetes Mellitus', WEB, vol. 18, no. Special Issue 02, pp. 311–331, Apr. 2021, doi: 10.14704/WEB/V18SI02/WEB18074.

[20] R. N. Patil, S. Rawandale, N. Rawandale, U. Rawandale, and S. Patil, 'An efficient stacking based NSGA-II approach for predicting type 2 diabetes', IJECE, vol. 13, no. 1, p. 1015, Feb. 2023, doi: 10.11591/ijece.v13i1.pp1015-1023.

[21] A. H. Syed and T. Khan, 'Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study', IEEE Access, vol. 8, pp. 199539–199561, 2020, doi: 10.1109/ACCESS.2020.3035026.

[22] D. Berrar, 'Cross-Validation', in Encyclopedia of Bioinformatics and Computational Biology, Elsevier, 2019, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.

[23] M. Maalouf, 'Logistic regression in data analysis: an overview', IJDATS, vol. 3, no. 3, p. 281, 2011, doi: 10.1504/IJDATS.2011.041335.

[24] T. Hastie, 'The Elements of Statistical Learning: Data Mining, Inference, and Prediction'. 01 2009. doi: 10.1007/978-0-387-84858-7.

[25] R. Tibshirani, 'Regression Shrinkage and Selection via the Lasso', Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.

[26] T. Chen and C. Guestrin, 'XGBoost: A Scalable Tree Boosting System', in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[27] C.-C. Chang, Y.-Z. Li, H.-C. Wu, and M.-H. Tseng, 'Melanoma Detection Using XGB Classifier Combined with Feature Extraction and K-Means SMOTE Techniques', Diagnostics, vol. 12, no. 7, p. 1747, Jul. 2022, doi: 10.3390/diagnostics12071747.

[28] Z. Zhao, H. Peng, C. Lan, Y. Zheng, L. Fang, and J. Li, 'Imbalance learning for the prediction of N6-Methylation sites in mRNAs', *BMC Genomics*, vol. 19, no. 1, p. 574, Dec. 2018, doi: 10.1186/s12864-018-4928-y.

[29] N. H. N. B. M. Shahri, S. B. S. Lai, M. B. Mohamad, H. A. B. A. Rahman, and A. B. Rambli, 'Comparing the Performance of AdaBoost, XGBoost, and Logistic Regression for Imbalanced Data', *ms*, vol. 9, no. 3, pp. 379–385, May 2021, doi: 10.13189/ms.2021.090320.

[30] F. Pedregosa et al., 'Scikit-learn: Machine Learning in Python', Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825–2830, 2011

[31] R. D. Joshi and C. K. Dhakal, 'Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches', *IJERPH*, vol. 18, no. 14, p. 7346, Jul. 2021, doi: 10.3390/ijerph18147346.

[32] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, 'Predicting Diabetes Mellitus With Machine Learning Techniques', *Front. Genet.*, vol. 9, p. 515, Nov. 2018, doi: 10.3389/fgene.2018.00515.

[33] L. Breiman, 'Random Forests', Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[34] S. Härner and D. Ekman, Comparing Ensemble Methods with Individual Classifiers in Machine Learning for Diabetes Detection. 2022.

[35] L. Qin, 'A Prediction Model of Diabetes Based on Ensemble Learning', in *Proceedings of the 2022 5th International Conference on Artificial Intelligence and Pattern Recognition*, Xiamen China: ACM, Sep. 2022, pp. 45–51. doi: 10.1145/3573942.3573949.

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