A Novel Hybrid Approach to Predict Diabetes Using Boruta and Genetic Algorithm

Kirti Kangra, Jaswinder Singh

Abstract—**Diabetes is a persistent metabolic condition that impacts millions of lives globally. To detect diabetes at a base level with more accuracy, feature selection approaches are important. The feature selection methods elevate the performance of diabetes prediction models by identifying the most important and informative features from a large number of potential features. The present study proposes a new method for combining the Boruta algorithm with a genetic algorithm to select features for predicting diabetes. The Boruta algorithm, a stable feature selection technique, uses random forest classifier to evaluate feature importance and filter out irrelevant features. In contrast, genetic algorithms refine the selected feature subset by using natural selection and genetic crossover mechanisms to optimize feature selection. To measure the performance of the suggested method, the PIMA Indians Diabetes Dataset is used and implemented, which is a recognized standard dataset for diabetes prediction. The Boruta algorithm was initially used to filter out important features, followed by the genetic algorithm to reduce and optimize the feature set. The efficacy of the model was assessed using multiple measures, including accuracy, precision, recall, and F1 score, on a distinct test set. Experiments demonstrate that the proposed hybrid Boruta-GA algorithm performed better than traditional feature selection methods in achieving high accuracy for diabetes prediction. The selected best subset of features included significant features that significantly contributed to determining the performance of predictions. The proposed model showed an accuracy of 99.13% for diabetes prediction.**

Index Terms—**Diabetes, Algorithms-Boruta and Genetic, SMOTE, NB, DT, KNN, PIMA, etc.**

I. Introduction

THE rapid socioeconomic growth has improved dietary structure and brought a positive impact on health. However, these lifestyle changes, which reduced physical activities and consumption of processed food, have led to the prevalence of chronic health challenges, and diabetes is one of the most common among them. Hyperglycemia, which includes rising blood glucose levels, arises from inadequate insulin secretion by the pancreas or the body's impaired utilization of insulin, ultimately resulting in diabetes. This condition can lead to serious problems over time, such as renal failure, cardiovascular disease, stroke, visual impairment, and amputation of limbs. Diabetes is often referred to as the "second killer" among modern diseases, with only cancer exceeding it in morbidity rates [1]. Diabetes was ranked among the top 10 leading causes of death in 2019 “Global Leading Cause of Death Survey” [2]. The “International Diabetes Federation” estimated that, based on projections, the number of adults with diabetes would hit 700 million by 2045 [3]. The annual expenditure on healthcare attributable to diabetes is around $760 billion. Due to the significant worldwide impact of this disease, a burgeoning necessity exists to design innovative techniques for early diagnosis and efficient management.

In this context, the advancement of machine learning (ML), particularly in disease diagnosis and medical image analysis, has revolutionized the extraction of valuable insights from medical data for chronic disease prediction. ML techniques enable the early identification of both diabetic and non-diabetic individuals, allowing healthcare professionals to prioritize high-risk patients during diagnosis while reducing the need for extensive human intervention. This prediction technique enables the implementation of early preventative treatments, thereby decreasing diabetes frequency, improving the standard of living, and promoting overall healthy life expectancy. It also relieves the economic and healthcare cost load for treating diabetes. These advantages are the strength of motivation behind our research in the area.

In ML, ensemble learning involves combining multiple individual classifiers in various ways to enhance classification accuracy and robustness [4]. The three primary types are Bagging, Boosting, and Stacking. Bagging randomly selects subsets from the training set to create sub-training sets for each base model [5]. It then aggregates predictions from all base models to generate the final predictions. Boosting is defined as the repeated process of training base models, such that each subsequent model assigns higher weights to instances previously misclassified by previous models, thereby giving more importance to these instances. Stacking attempts to correct errors by iteratively aggregating the output of all the base models using a weighted linear technique [6]. Stacking, on the other hand, aggregates several base models and a meta-model. Predictions from the base models are used as input to the meta-model, which is then trained to produce the final classification output.

Ensemble classifiers are better than single classifiers in classification. Developing a classification model that achieves high robustness and accuracy while maintaining efficient time and space complexity remains a challenging goal. However, in practical classification scenarios, the performance of a classifier is significantly influenced by dataset errors such as outliers, extremes, and noisy data, impacting classification outcomes. Single classifiers are particularly vulnerable to degraded performance when encountering noisy data, leading to a decline in accuracy. In contrast, ensemble classifiers assign varying weights based

Manuscript received Feb 27, 2025; revised August 2, 2025.

Kirti Kangra is a Research Scholar at Guru Jambheshwar University of Science and Technology, Department of Computer Science and Engineering, Hisar, Haryana, India (e-mail: kirtikangra98@gmail.com).

Jaswinder Singh is a professor at Guru Jambheshwar University of Science and Technology, Department of Computer Science and Engineering, Hisar, Haryana, India (e-mail: jaswinder_singh_2k@rediffmail.com).

on voting, enabling them to reclassify misidentified data and exhibit superior adaptability to noisy data. Consequently, this study adopted ensemble classifiers due to their ability to address these challenges. The classification of diabetic patients differs from other datasets due to diabetes being a prevalent chronic disease. Medical datasets, too, differ from others in that disease diagnosis is a complicated, multi-faceted process involving economic, physical, and psychological factors, particularly in chronic diseases. Misdiagnosing a disease could have fatal repercussions for the patient. Hence, in disease diagnosis, particularly for conditions like diabetes, selecting a classifier holds greater importance, prioritizing accuracy as a crucial aspect.

A. Feature Selection

Feature selection is a significant aspect of data analysis and ML. It aims to extract the most pertinent and instructive features from an extensive dataset. It enhances ML model performance by diminishing dimensionality, alleviating overfitting risks, accelerating training, and leveraging optimization techniques.

Feature selection methods can be divided into three major types: Filter, Wrapper, and Embedded approaches. The Filter methods assess the importance of individual features using statistical metrics, such as variance thresholding, correlation analysis, and the Chi-squared test [7]. The Wrapper method evaluates different feature subsets by training and testing ML models to determine the combination that yields the best performance [8]. It includes "recursive feature elimination (RFE), forward selection, and backward elimination". Embedded methods incorporate feature selection firmly into the model training process, including techniques such as "sequential feature selection, domain knowledge, tree-based models, dimensionality reduction, and L1 regularization (Lasso)" [9]. In addition to these traditional approaches, metaheuristic algorithms are widely used for feature selection, as they efficiently explore the search space to identify optimal or near-optimal feature subsets [10]. Examples of such techniques include "simulated annealing, genetic algorithms, and particle swarm optimization", which evaluate different feature combinations based on an objective function or fitness criterion.

Furthermore, hybrid approaches integrate multiple feature selection techniques to leverage their strengths while mitigating their limitations. For instance, a filter method may be used initially to select relevant features, followed by a wrapper method for fine-tuning the final subset. This synergy enhances selection efficiency and improves model performance.

B. Objective

To diagnose diabetes at an early stage through the examination of different contributing factors is the focus of this study. To determine the most relevant features, it has utilized a combination of the Boruta algorithm and a genetic algorithm for feature selection. This hybrid strategy has been demonstrated to improve the accuracy of diabetes prediction models. To correct any class imbalance in the selected data, the "Synthetic Minority Over-sampling Technique (SMOTE)" is implemented, which is well-suited to enhance model performance on imbalanced data. The optimized feature set is utilized to train a range of ML classifiers.

The present study is outlined in the following order: Section II outlines a summary of the prior works on diabetes prediction. Section III gives a presentation of the proposed model and methodology, including the Boruta algorithm, genetic algorithm, SMOTE, and the classification methods employed. Section IV offers a comprehensive description of the experiment, encompassing a description of the dataset and its features. It outlines the stepwise procedures followed and provides the parameters established for the experimental setup. The results obtained through experimentation are presented in the next section and compared with other models. Consequently, Section VI presents the outcomes and future concerns of this study.

II. Previous Work

Diabetes is a serious and chronic disease that affects the human body in many different ways. Several research works have been done to identify and predict diabetes with ML techniques to extract features.

Zaiheng Zhang et al. (2024) introduced the AHDHS-Stacking ensemble learning system for the diagnosis and classification of diabetes mellitus. It employs the stacking method and Harmony Search (HS) algorithm, which combines two essential steps: feature selection and base-learner ensemble optimization. The trial used the "Pima Indians Diabetes (PID) and the Chinese and Western Medicine Diabetes (CWMD)" datasets. The study achieved outstanding performance measures on the PID dataset, including 93.25% F-measure, 84.79% MCC (Matthews Correlation Coefficient), 93.09% accuracy, 93.22 % precision, and 91.60 % recall [11].

Hongfang Zhou et al. (2023) concentrated on Boruta feature selection to retrieve significant features from datasets. The researchers utilized an ensemble learning strategy for classification and the K-Means++ technique for unsupervised clustering. The study yielded an astounding 98% model accuracy rate using the PID dataset [12]. Patil et al. [13] introduced "NSGA-II-Stacking", a stacking-based evolutionary ensemble learning framework for type-2 diabetes mellitus prediction. Developed in MATLAB and applied to the PID dataset (with missing values imputed using median imputation), their approach employed KNN as the meta classifier and a multi-objective optimization algorithm as the base learner, with an F-measure of 88.5%, a ROC value of 85.9, sensitivity of 96.1%, specificity of 79.9%, and overall accuracy of 83.8%. Su et al.. used the PID dataset, XGBoost, LightGBM, Neural Network, and LR algorithms and found the result as federated learning models. The study illustrates that it could be utilized better while selecting data of the patients from other organisations, which in turn may produce a more accurate and consistent risk prediction for Diabetes Mellitus [14]. In addition, Pooja Yadav et al. [15] developed a grid search-based improved grey wolf method. Boruta for feature selection and the SMOTE method for dataset balancing in the study. It evaluated the prediction model's performance with a focus on the Stacking Classifier, and the results showed that the Proposed Model had the highest F1-score of 98.84% on the PID dataset. However, Ayşe Doğru et al. (2023) [16] presented a novel super ensemble learning model to promote diabetes mellitus early diagnosis. This model combines predictions from different

ML methods by cross-validation. It consists of four base learners (LR, DT, RF, and gradient boosting) plus a meta-learner, which shows better accuracy in identifying diabetes mellitus. Out of five methods, chi-square was found to be the best feature selection strategy; Grid Search was then used to adjust the hyperparameters. Outstanding accuracy rates of 99.6% for the Sylhet Diabetes dataset, 92% for the PID dataset, and 98% for the “diabetes 130-US hospitals” dataset were attained by the super learner. Hairani and Dadang Priyanto (2023) have developed the SMOTE-ENN methodology and used the PID dataset to predict diabetes and improve the performance of RF and Support Vector Machine (SVM) classification algorithms. By using class balancing and removing noisy data close to class boundaries to reduce dataset imbalance, the RF algorithm with SMOTE-ENN surpassed SVM with an accuracy of 95.8% [17]. Moreover, Dipesh Kumar et al. proposed a fog-based diabetes prediction model utilizing patient data sensed from remote sensors [18]. A hybrid technique known as ANFIS-PSO-WOA was utilized at the cloud layer for the detection of diabetes, and real-time data processing at the device level was facilitated through fog computing. Testing with UCI repository data showed that the proposed method resulted in a very satisfactory prediction accuracy of 92%.

In 2023, Chetan Nimba Aher and Ajay Kumar Jena proposed the “Improved Invasive Weed Bird Swarm Optimisation Algorithm (IWBSOA)” for predicting diabetes. Their approach combines the “Bird Swarm Algorithm (BSA)” with an enhanced “Invasive Weed Optimisation (IWO)” using both Recurrent Neural Network (RNN) and SVM classifiers. The hybrid deep learning model achieved remarkable performance metrics—96.19% accuracy, 97.11% sensitivity, 94.39% specificity, and an MSE of 0.1887 [19].Employing a dataset from the Gene Expression Omnibus database, Rajagopal et al. created a new hybrid model that combines an “artificial neural network (ANN) “and a “genetic algorithm” for the prediction of diabetes. This method efficiently analyzes the effect of every variable by prioritizing the most significant features, reaching an 80% prediction rate on the PID dataset [20]. Shamreen Ahamed and Sumeet Arya (2022) conducted a series of experiments using seven different ML techniques on the PID diabetes dataset. Out of these, the LGBM algorithm performed the best with a 95.2% accuracy [21]. Selim Buyrukoglu and Ayhan Akbas integrated correlation heatmaps with sequential forward selection (SFS) in 2022 to find the best subsets of features. They later used SVM, RF, and ANN classifiers using the selected features, with ANN-based hybrid feature selection achieving a whopping 99.1% accuracy on the Sylhet Diabetes dataset [22]. Also in 2022, Altyeb Altaher Taha and Sharaf Jameel Malebary [23] introduced a novel ensemble learning method for type-2 diabetes prediction. Their approach, which integrated fuzzy clustering with logistic regression (LR) in a hybrid meta-classifier, attained accuracies of 99.00% for the PID dataset and 95.20% for the SDD dataset. Reza Ghabousian et al. [24] suggested a new method that combined fuzzy inference systems and particle swarm optimisation metaheuristics. By integrating the particle swarm algorithm in binary form using fuzzy systems, their method achieved an impressive classification accuracy of 95.47%. Gizen Mutlu and Çigdem Inan Acıcreated a parallel-hybrid model based on SVM, “Sequential Minimal Optimisation (SMO), and Stochastic Gradient Descent (SGD)” to predict diabetes with an overall accuracy rate of 87% [25]. Michael Onyema Edeh et al. compared several algorithms on various datasets. They observed that the RF algorithm had the best accuracy (97.6%) on the Frankfurt Hospital database in Germany, whereas the SVM algorithm attained 83.1% accuracy on the PID dataset [26]. Xiaohua Li, Jusheng Zhang, and Fatemeh Safara (2021) suggested an integrated strategy that combined feature selection, classification, and preprocessing employing K-means clustering with many feature selection algorithms to attain 91.65% accuracy on the PID dataset [27]. In another 2021 study, Satish Kumar Kalagotla, Suryakanth V. Gangashetty, and Kanuri Giridhar implemented a three-phase strategy. They started with correlation-based feature selection, followed by AdaBoost for classification, and then designed a bespoke stacking approach using MLP, SVM, and LR specifically for the chosen features. This method predicted diabetes with an impressive 97.4% accuracy [28]. N. Kanimozhi and G. Singaravel suggested a stacking-based integrated “kernel extreme learning machine (KELM)” model to identify high-risk individuals for type II diabetes. Using “Artificial Fish Swarm Optimization-Hybrid Particle Swarm Optimization” (HAFPSO) to minimize kernel complexity and maximize accuracy, their model attained a value of 98.5% [29]. M G Dinesh and D. Prabha [30] used kernel principal component analysis for feature reduction along with a genetic algorithm for feature selection. Likewise, C. Mallika and S. Selvamuthukumaran [31] proposed an effective diabetes diagnosis technique that combined SVM classification with optimization using the “Crow Search Algorithm (CSA) and Binary Grey Wolf Optimizer (BGWO)”, evaluating their method on the PID dataset. Shirina Samreen developed an early diabetes diagnosis technique based on a machine learning pipeline. The researcher used an ANOVA filter, “Crow Search Optimisation, and Singular Value Decomposition” as feature selection methods, followed by stacking ensemble of different classifiers (AdaBoost, GradientBoost, LR, K-NN, DT, SVM, RF, and Naive Bayes), which yielded high accuracy of 98.4% with the minimal feature set [32]. Saloni et al. proposed an ensemble soft voting classifier for binary classification by integrating RF, LR, and NB. Their experimental comparison—also with other ensemble methods and standalone classifiers like AdaBoost, SVM, and CatBoost—resulted in an accuracy of 79.04%, a precision of 73.48%, a recall of 71.45%, and an F1 score of 80.6 on the PID dataset [33]. Rajendra et al. compared linear regression and ensemble learning approaches on the PID dataset and reported that LR performed exceptionally well in building predictive models. Their work highlighted the significant contributions of data pretreatment, feature selection, and integration methods in improving model accuracy [34]. In a similar investigation, A. Singh, A. Dhillon, and N. Kumar integrated different ML methodologies (“XGBoost, RF, SVM, neural networks, and DT”) with the eDiaPredic ensemble model to predict diabetes. Tested on various measures—such as sensitivity, accuracy, precision, and the Gini Index—their method was found to have a 95% accuracy level on the PID dataset [35]. Islam et al. presented two novel feature selection techniques using wavelet decomposition and the fractional derivative for diabetes prediction. They preprocessed “oral glucose tolerance test (OGTT)” data by imputing missing values with the arithmetic mean and then used classifiers like “SVM, NB, RF, AdaBoost, and Bagging”. From the San Antonio Heart Study data, their system reported 95.94% accuracy[36]. Rajendar et al. [37] used different ML methods, such as “DT, LR, RF, and SVM,” for the PID dataset to predict diabetes, with the highest accuracy being obtained by SVM for predicting diabetes risk. Likewise, Tripathi and Kumar [38] compared four ML models “(LDA, KNN, SVM, and RF”) on the UCI-sourced PID dataset for the prediction of early-stage diabetes, and the best performance was reported by RF with an accuracy of 87.66%.

In general, these studies suggest that although much has been achieved, still more research needs to be conducted to enhance the accuracy of diabetes prediction. Building a strong classifier—or collection of classifiers—that reduces error rates is still important for being able to distinguish reliably between patients with and without diabetes.

III. Proposed Method

This study suggests a diabetes predictive model incorporating feature selection procedures based on Boruta and Genetic algorithms, with performance improvement by ensemble learning procedures based on stacking (see Fig. 1). Missing values in the dataset are initially handled by replacing them with the mean. Subsequently, the feature selection procedure, based on both Boruta and Genetic algorithms, removes the features that are extraneous and determines the most important ones for the diagnosis of diabetes. The model uses a stacking ensemble in which “Naive Bayes (NB) and Decision Tree (DT) are the base models, and K-Nearest Neighbors (KNN)” is used as the meta-model. In this procedure, the original data is preprocessed first and then passed through the feature selection algorithms to identify the most appropriate features, which are subsequently normalized. These normalized features are then passed into the stacking classifiers for final classification, with the meta-model combining the predictions of the base models to decide if a patient has diabetes. Below is an overview of the proposed method’s algorithm:

Step 1: Load the PID dataset and replace missing values with the mean.

Step 2. Rank the dataset features based on the support function as a selector for the Boruta algorithm.

Step 3: After that, from the selected features, determine a suitable subset via a Genetic algorithm using a fitness function.

Step 4: In the proposed method, after feature selection, features were normalized, and the SMOTE technique was applied to eliminate any types of imbalances.

Step 5: In classification stacking ensemble learning, at the 0 level, DT and NB were employed, and KNN at level 1 or as a meta classifier was employed.

Step 6: Performance evaluation was executed utilizing multiple metrics, including accuracy, precision, recall, ROC curve, Kappa value, RSE, RRSE, MAE, RMSE, MCC, and MSTSS to evaluate both predictive quality and error rates.

A. Background

ML, a swiftly evolving technological domain, has become one of the most effective approaches for tackling intricate and multifarious issues. The implementation of ML techniques in healthcare is accelerating due to the automatic pattern identification procedures associated with this field. This research utilized various ML techniques for the proposed diabetes prediction model. This section of the study briefly introduces how each technique works.

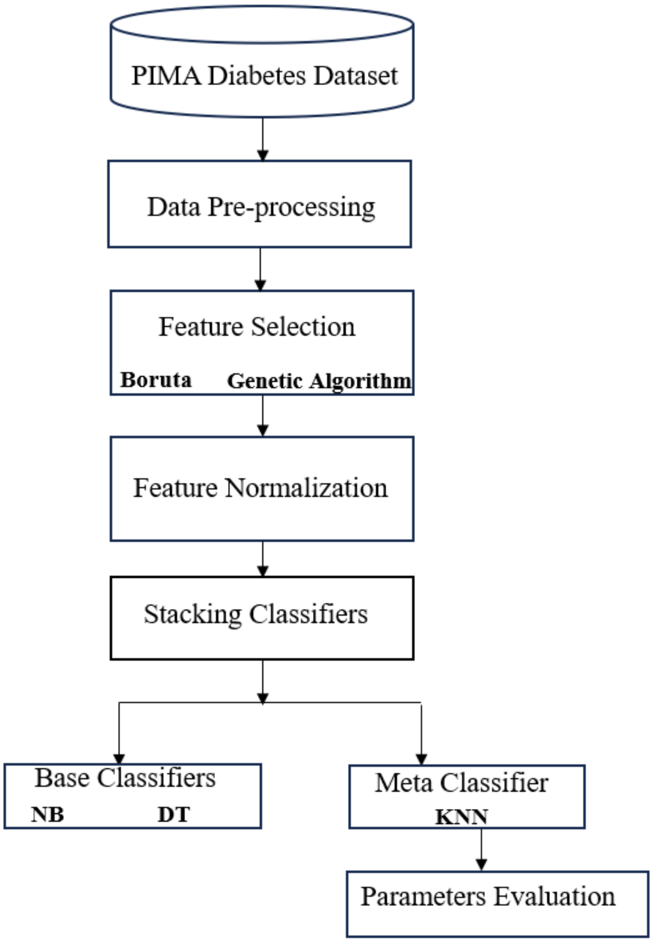

Fig. 1. Methodology for the Proposed Model

a. Boruta Algorithm

Boruta algorithm is a feature selection technique that is used for discovering relevant features in noisy or complex structured data sets. It operates in a manner that compares the feature importance of each feature to that of shadow features, which are generated randomly. Therefore, it can identify the features that are statistically significant for predicting the target variable. The working steps are as follows:

1. The model first calculates the feature importance scores with a Random Forest (RF) classifier as a baseline for comparison.

2. Then, it creates more random “shadow” features by permuting the original features, which act as a baseline for assessing the importance of the real features.

3. Each original feature is evaluated by comparing its importance to that of its shadow features. If it scores higher than its shadow counterparts based on a set significance threshold, it is deemed important and kept.

4. The process is then repeated iteratively with the algorithm continually evaluating feature importance and removing those that are not statistically significant until only the most significant features are left.

5. Lastly, the algorithm returns the subset of features that are deemed statistically significant for predicting the target variable.

By selecting the most significant features first, Boruta offers an automated and reliable feature selection technique that maximizes both interpretability and ML model performance. It determines the most important features of the dataset for the target variable by combining support and ranking metrics. This study used only the support function.

b. Genetic Algorithm

Genetic Algorithm (GA) is a metaheuristic approach for searching, based on Darwinian evolution principles, i.e., natural selection [39]. The algorithm picks the fitness individuals for reproduction in the next generation, essentially mimicking the process of natural selection. The following outlines how GA operates:

1. First, a random population is generated.
2. The algorithm produces a sequence of new populations. The subsequent population uses the past population to produce. To create the new population, the following actions should be taken:
    a. The algorithm calculates the fitness score.
    b. The algorithm identifies individuals based on their fitness score to be called parents.
    c. The least fit members of the present population are chosen as elites and transferred to the succeeding population.
    d. Offspring are generated by the algorithm from their parents. Two methods are used to produce offspring: applying random modifications to a single parent (known as a mutation) or combining the vector components of two parents (known as crossover). To form the succeeding generation, the algorithm replaces the current individuals with their offspring.
3. The algorithm terminates when either the time limit or the fitness limit, such as the specified number of generations, is reached.

c. SMOTE

It addresses the issue of class imbalance in ML datasets. The imbalanced data sets can create biased models with poor performance over minority classes. Samples between instances already exist to generate synthetic samples for the minority class to counteract this issue [40]. Specifically, the present method determines the closest neighbors of instances randomly chosen from the minority class. Then, it generates synthetic samples by creating new instances along the line segments joining the selected instance with its nearest neighbors [41]. In this way, this method for ML contributes to balancing the distribution across classes by effectively increasing the representation of the minority class in the dataset. Through the use of more diverse examples, this approach enables classifiers to learn better and achieve higher accuracy for both majority and minority classes [42]. Consequently, SMOTE is effective in addressing class imbalance while retaining model performance. It is crucial to carefully examine the results and be aware of any potential downsides, such as the introduction of noise into the dataset.

d. DT

DT is a simple, intuitive algorithm implemented for “classification as well as regression” problems. The algorithm recursively divides the dataset into subsets based on the features to best distinguish among the data that fall into disparate groups or those that predict values of a different kind [43]. In each decision tree node, a feature is represented by each node, and the branches present the potential decision or outcome against that feature. This division process repeats until a certain depth or termination criterion is met. Due to its linear structure, decision trees are simple to comprehend and visualizable, successfully incorporating both numerical and categorical input.

e. RF

RF is an ensemble learning algorithm that creates many decision trees using randomly selected subsets of the training data and attributes. For classification, the prediction is achieved using a majority vote across the individual trees. The total tree production yields the final prediction. RF has become renowned for its resilience to overfitting, robustness, and capacity to manage huge datasets.

f. NB

NB is a probabilistic classification technique based on the Bayes theorem and the feature autonomy assumption. It assumes that a feature's presence in a class is irrelevant to other features; it computes the possibility of a data point corresponding to a specific class based on its features. NB can be exceptionally efficient in many real-world circumstances, despite its “naïve” assumption of feature independence [44]. It is especially helpful for text classification and spam filtering.

g. KNN

KNN is a particularly simple and easy-to-understand technique that is used for regression as well as classification. It determines a test data point's closest neighbours from the training dataset that is based on a selected distance metric (such as Euclidean distance) in the feature space [45]. This process operates on the proximity principle. In the KNN algorithm, the test point is evaluated based on the mean or majority vote of the labels from the k nearest neighbors, where k is a user-defined value. The accuracy of a KNN model depends on the selection of a proper distance function and the optimal selection of k.

h. Ensemble learning

In the stacking ensemble learning technique, a single meta-classifier merges multiple classification models. It uses many base models and aggregates their results to train a meta-model that produces the final result through continuous training. The base and meta-models for stacking in this study are NB, DT, and KNN, in that order. The steps for the stacking method are shown in Algorithm 1 below.

Input: Selected features $SF=\{(x_i,y_i)\}^n_{i=1}$, Base Model $BM_1,BM_2,BM_3\ldots,BM_n$ Meta Model MM

Output: Parameter evaluation of the classifiers after stacking

Steps:

1. Split the selected features SF into training and testing data
   $SF_{train}=\{(x_i,y_i)\}_j^{\,i}=1,\ SF_{test}=\{(x_i,y_i)\}_k^{\,i}=1$
   Where x is the feature class attribute and y is the output class attribute
2. Use SF_{train} data to train Base Model BM_1, BM_2, $BM_3 \ldots BM_n$
3. Now, construct a new train dataset to train the meta-model.
4. SF_{test} is sequentially fed into the trained base model, and a new test dataset is generated.
5. Output the classification results of the stacking model using a new test dataset by feeding it into a trained meta-model.

IV. Experiment

This section of the paper provides information regarding the material and methods followed by the study.

A. Datasets

This study used the PIMA dataset related to Diabetes. The dataset of type 2 diabetes comprises 768 records and nine important variables [46].

B. Pre-processing

The dataset underwent a comprehensive preprocessing phase to ensure data integrity and suitability for analysis. One crucial step involved handling missing values, which were replaced with their respective mean values using the following (1) [47]. The goal of this procedure was to minimise the effect of missing data on further evaluations without sacrificing the dataset's structural integrity. The dataset's statistical features were preserved by imputing missing values using the computed means, enabling a more thorough and reliable analysis. This preprocessing stage ensured the correctness and dependability of the outcomes from the following analytical techniques by enabling the modeling and analysis processes to be carried out on a fuller and more representative dataset.

$$Dataset(\text{data})=mean(\text{data})\ if\, value=null \quad \text{else data} \qquad (1)$$

C. Procedure to select features

In this study, four features: "Plasma glucose concentration at 2 hours in an oral glucose tolerance test, Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function, and Age (years)" were chosen using the Boruta and Genetic algorithm (see algorithm 2), for the "PIMA Indian diabetes" dataset. The selected features associated with the data were saved and then analyzed further.

Algorithm 2: For Feature Selection

Input: PIMA Indian Diabetes Dataset D with missing values and irrelevant features

Output: Selected Features (SF)

```
    For (i=1 to n)
     If (Feature values==missing values/ zero)
          Replace with mean using Eq. (1)
      Else
           Feature value = Value
Apply the Boruta algorithm for feature selection
            Using Support Eq.(2)
       After that, from Selected Features ($S_i$) using Boruta, apply
the Genetic Algorithm to find a proper subset of features using:
          For each selected feature, $S_i$ to P do // P=Population
           Evaluate the Fitness function
   While iteration number < n // n= max number of generations
             Select= SelectBst(i);
             If Select then // using tournament selection
               If Cross-over, then // two-point crossover
                 Choose two parents, $i_a$ and $i_b$
                  Produce offspring $i_c$= cross-over
              Else
                 Choose one individual
                  Produce offspring by Mutate($i_c$)
            Terminate
               Evaluate the fitness value of $i_c$;
               Replace with the least fitness value feature
```

After this procedure selected feature subset will be there, which has a high fitness function.

D. Feature Normalization

The study employed the Min Max scaling and SMOTE techniques to normalize the data and eliminate any imbalance in this experiment. The data was split into a 70:30 ratio, and validation was performed using 10 cross-folds.

Algorithm 3: For Proposed Method

Input: Selected Features

Output: Parameter Evaluation

Start:

1. Feature Scaling using Min Max
2. To remove any kind of imbalance using SMOTE
3. Refined features are fed into the stacking model (using Algorithm 1)
4. Parameter Evaluation using Accuracy, Recall, Precision, Kappa value, MCC (see the experimental result for equations used)
5. Error rate calculation using MSE, RMSE, RAE, etc.

E. Experimental Setup

Through the preprocessing, the study determines the most relevant features from the dataset via feature selection techniques. After normalizing the processed data, the chosen features were used as input for further processing. The above experiment was conducted using a Jupyter notebook on a system equipped with an "AMD Ryzen 5 5500U with Radeon Graphics and 16 GB RAM under x64 bit Windows 11 operating system".

F. Experimental Results

The proposed model was evaluated using test data from the PID dataset. The dataset's performance has been assessed using various evaluation metrics, including F-measure, recall, precision, and classification accuracy. Applying the Boruta and Genetic algorithms integrated feature selection strategy with the stacking classifiers DT, NB, and KNN resulted in an accuracy rate of 99.13% for a 70:30 split. The suggested model can help medical professionals make better selections by utilising features that have been extracted. The different properties of the PID dataset are used to execute various algorithms. The parameter values used for each method are mentioned in Table I below.

TABLE I
PARAMETER VALUES

Boruta Algorithm	

RF	n_estimators=100,
	random_state=42

Genetic Algorithm	
Population size	100
max_generations	50
Crossover Rate	0.8
Mutation Rate	0.3
Type of Selection	Tournament Selection
Fitness function	Accuracy
Cross over	Two Point
Model	Parameter

DT	max_depth=16,
	criterion='entropy',
	random_state=42

KNN	n_neighbors=5,
	algorithm='auto'

NB	GaussianNB

V. RESULTS

This section of the study provides results regarding different computations done on the selected datasets.

A. Evaluation Parameters

This study evaluates the efficacy of the applied model through multiple metrics, including "Matthews Correlation Coefficient (MCC), F1-score, accuracy, recall, and precision. The classifier's classification results are displayed in a matrix called the confusion matrix. True-negative (TN), false-positive (FP), false-negative (FN), and true-positive (TP)" are all included in this classification. Evaluation metrics developed by TP, FP, FN, and TN facilitate the process of evaluating the performance of the implemented model. When examining the quality of a binary classification model, the MCC metric is utilised. The scale has values ranging from -1 to 1, where 1 represents an accurate prediction, 0 represents a result that is no better than chance, and -1 represents total disagreement.

a. Model Evaluation (70:30)

Table II comparisons revealed that the introduced model exhibited 0.1–13.1% increased accuracy, 0.4–12.9% elevated recall, 0.9–24.7% enhanced F1-score, 0.9–52.4% higher Kappa coefficient, 2.4–29.2% improved precision, and 2.2–64.3% greater MCC coefficient compared to the alternative methods. Consequently, the implemented model performs better and beats the current prediction models in every evaluation metric. To determine the first value, deduct the introduced model's value from the highest value in the row. In addition, to determine the second value, compute the mean value of all the models, excluding the proposed model. Moreover, subtract the mean value from the proposed model value, and divide the final subtraction result by the proposed model value.

b. 10 Cross-Validation

Based on the mean values presented in Table III for comparison, the suggested model exhibited 0.2–22.8% increased accuracy, 0.3–15.4% elevated recall, 0.7–26.5% enhanced F1 score, 1.8–61.7% higher Kappa coefficient, 0.6–14.5% improved precision, and 2.2–63.5% greater MCC coefficient evaluated against other models. Therefore, the proposed model is better and outperforms other prediction models on all the evaluation metrics.

B. Comparison in the raw dataset

This segment utilizes the unprocessed PID dataset, which contains noisy outliers and missing values. The data is utilized to validate the model suggested in this research and is evaluated against some baseline ML models. The forthcoming tables provide a clear description of the experimental results.

a. Model Evaluation

In Table IV, the performance of the introduced model is evaluated against other traditional models using the original PID dataset split into a 70-30 ratio. By surpassing the other models by 3.6-20% in accuracy, 7.2-16.7% in precision, 3.7-17.6% in MCC, and 2.7-15.7% in kappa value, the introduced model exhibits remarkable performance across several evaluation metrics.

b. 10 Cross-Validation

In Table V, the proposed model demonstrates a higher accuracy ranging from 0.6-3.1%, precision improvement of 1.8-10.0%, F1- score increase of 0.1-3.2%, significant MCC enhancement ranging from 12.9- 29.5%, and elevated kappa values ranging from 0.8-6.9% and compared to the baseline models.

c. Standard deviation test

Table VI compares the performance of several ML models with the proposed model using various metrics, such as MCC, Kappa value, F1-score, accuracy, recall, and precision. To show the variety of performance throughout several iterations or cross-validation folds, each model is evaluated using its mean performance as well as the standard deviation. The proposed model performs well across several metrics. It is essential to consider both mean performance and variability when evaluating the robustness and reliability of the ML model.

C. Performance Evaluation using Different Datasets

To assess the current performance of this model and validate its applicability and dependability, the study conducted additional testing on a new dataset of diabetic patients, which was received from the Hospital Frankfurt diabetes dataset [48]. The chosen features are 'Age', 'Glucose', 'Skin Thickness', 'Insulin', 'BMI', 'Diabetes Pedigree Function', and 'Diabetes'.

a. Model Evaluation (70:30)

The new model surpasses other ML models in terms of various evaluation parameters. Specifically, it exhibits a significant improvement of 1.3- 14.7% in accuracy, 0.6-14.9% in recall, 1.8-23.6% in precision, 1.3-19.3% in F-1 score, 0.6-30.4% in MCC, and 0.8-30.9% in kappa value compared to the alternative models (see Table VII).

b. 10 Cross-Validation

The new model an accuracy improvement ranging from 2.1-16.1%, recall enhancement ranging from 3.3-17.2%, precision increase ranging from 0.7% -17.1%, F-1 score elevation ranging from 1.8-17.0%, MCC improvement ranging from 4.3-34.3%, and kappa value enhancement ranging from 4.4-34.6% as shown in the Table (VIII).

D. Comparison in the raw dataset

The raw Frankfurt diabetes dataset, which contains missing values and noisy outliers, was used in this section. This dataset will be utilised in the investigation to assess the proposed model and contrast it with other traditional ML models. The experimental results are broken down in depth in the tables below.

a. Model Evaluation (70:30)

In Table IX, the performance of the DT classifier is comparable to that of the new model in this study. However, the introduced model proves to be more effective than the other models overall. Specifically, the new model demonstrates higher accuracy, ranging from 14.8- 13.4%, recall ranging from 29.1-28.3%, precision, ranging from 13.9-13.8%, F-1 score ranging from 22.9-21.8%, MCC ranging from 35.9-31.9%, and kappa value ranging from 33.8-32.6%.

b. 10 Cross-Validation

In Table X, the introduced model demonstrates a wide range of performance metrics. Specifically, it shows an improvement in accuracy ranging from 1.6- 37.5%, an increase in recall ranging from 6- 35.6%, a precision improvement ranging from 18.8-19.4%, an enhancement in F-1 score ranging from 3- 28.8%, a significant increase in MCC ranging from 7- 43.0%, and an elevation in kappa ranging from 4-41.7%. The results reinforce the model's robustness and efficacy across multiple evaluation metrics.

c. Standard deviation test

The performance of several ML models and the introduced model is assessed in the above table based on some measures, such as MCC, Kappa value, F-1 score, accuracy, recall, and precision. The performance of each model and the standard deviation across multiple iterations or cross-validation folds are presented. From the results, the study observes that the introduced model for the 'Frankfurt' dataset generally exhibits competitive performance across most metrics, with relatively low standard deviations evaluated against other models. This suggests that the performance of the proposed model is consistent across different evaluations (see Table XI).

E. ROC Curve

Plotting TPR (Sensitivity) versus FPR (1 - Specificity) at different threshold values forms the ROC curve. A distinct threshold setting is represented by each point on the curve. Better performance is indicated by a higher curve that is closer to the top-left corner of the plot; this suggests that the model maintains a low FPR across a range of threshold values while achieving greater TPR. The ROC curves for both datasets are shown in Figures 2 & 3.

F. Error Rates

Error rates are a collection of variables used in ML models to assess how well predictions match actual or predicted outcomes. Error rate reduction is the goal for optimal outcomes. These metrics provide insight into the size, precision, and variability of prediction errors, among other aspects of model performance. The equations in Table XII were utilized in this investigation to calculate the error rates. The "Root Mean Squared Error (RMSE)" is a commonly used method to evaluate model error in statistical data prediction. Its values lie between 0.0 and 0.5, implying strong prediction accuracy. "Relative Root Squared Error (RRSE)" is a crucial indicator for evaluating the performance of models; lower values indicate better performance. Relative Squared Error (RSE) is used to anticipate the target's mean by comparing the squared error of a regression model to that of a basic baseline model. A low RSE indicates good model performance, while values near one suggest no discernible improvement over the baseline model. The "Mean Absolute Error (MAE)", quantifies the average absolute deviation across predicted values and actual values, serving as a commonly used statistic in regression models. Lower MAE values are indicative of higher model accuracy[49]. The "Mean Squared Total Sum of Squares (MSTSS) quantifies the total variation in a dataset by summing the squared deviations of each data point from the overall mean and subsequently dividing this sum by the total number of data points". Depending on the requirements, it is advised to give priority to the classifier with the lowest priority to the classifier with the lowest RRSE or MAE, and RMSE when choosing one based on these parameters. For better classifier predictions, low RRSE values are especially required. Table XIII illustrates the results for error rates of the two datasets.

Overall, the interpretation of these metrics suggests that the model is performing very well with high accuracy, precision, recall, relatively low errors, and discrepancies compared to baseline models with the latest research.

VI. Discussion

This study emphasizes the significance of utilizing techniques for selecting features to identify the most pertinent features for diabetes diagnosis. Identifying features that correspond to doctors' diagnostic criteria is more important than focusing on optimizing performance measures. In real-world scenarios, features that might just improve performance without aiding in an accurate diagnosis are considered less significant. Conversely, even while some diagnostically important features result in unsatisfactory predictive results, they are still significant since they enable doctors to make knowledgeable selections.

Therefore, this study utilizes the Boruta and Genetic algorithms for selecting features, after which Min-Max scaling and the SMOTE method are applied to normalize the chosen features and correct any class imbalance. Our model is further improved with ensemble learning and parameter tuning. Experimentation on two datasets, combined with careful comparison with other approaches, shows minimal error rates in all major experimental measures. Interestingly, feature selection is done before normalization to highlight significance, decrease complexity, enhance interpretability, avoid overfitting, and handle imbalance more effectively.

One of the major limitations of our research is the limited dataset size, which could impact model training robustness. In future work, efforts will be made to acquire a broader and more realistic diabetes dataset to further reduce the threat of undertrained models.

VII. Conclusion

Thus, the combination of Boruta and Genetic Algorithms presents a promising feature selection technique for diabetes prediction. Experiments on the PID diabetes dataset attained a remarkable accuracy of 99.1% using tenfold cross-validation. This research utilized a stacking ensemble for classification with NB and DT as base models at level 0 and KNN as the meta-model at level 1. Compared to other models, this solution presented better results. Validation of the Hospital Frankfurt dataset also attested to the strength and reliability of our model in diabetes detection. The present study not only upturns prediction accuracy but also provides information about the most instructive features that were involved in the diagnosis of diabetes. This work improves feature selection methods for diabetes prediction, and it may be applied to various healthcare contexts.

References

[1] "The top 10 causes of death." https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed Dec. 20, 2023).

[2] International Diabetes Federation, "Facts & figures." https://idf.org/about-diabetes/facts-figures/ (accessed Jul. 04, 2023).

[3] H. Salem, M. Y. Shams, O. M. Elzeki, M. A. Elfattah, J. F. Al-amri, and S. Elnazer, "Fine-Tuning Fuzzy KNN Classifier Based on Uncertainty Membership for the Medical Diagnosis of Diabetes," *Appl. Sci.*, vol. 12, no. 3, pp. 1–26, 2022, doi: 10.3390/app12030950.

[4] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/9930985.

[5] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021, doi: 10.1016/j.icte.2021.02.004.

[6] G. T. Reddy *et al.*, "An Ensemble based Machine Learning model for Diabetic Retinopathy Classification," in *International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE 2020*, 2020, pp. 1–6, doi: 10.1109/ic-ETITE47903.2020.235.

[7] R. Parthiban *et al.*, "PROGNOSIS OF CHRONIC KIDNEY DISEASE (CKD) USING HYBRID FILTER WRAPPER EMBEDDED," *Eur. J. Mol. Clin. Med.*, vol. 07, no. 09, pp. 2511–2530, 2020.

[8] M. Manonmani and S. Balakrinshnan, "An Ensemble Feature Selection Method for Prediction of CKD," in *2020 International Conference on Computer Communication and Informatics(ICCCI-2020), Jan. 22-24,2020,Coimbatore,India*, 2020, pp. 20–25, doi: 10.1109/ICCCI48352.2020.9104137.

[9] V. Kumar, J. K. Chhabra, and D. Kumar, "Parameter adaptive harmony search algorithm for unimodal and multimodal optimization problems," *J. Comput. Sci.*, vol. 5, no. 2, pp. 144–155, 2014, doi: 10.1016/j.jocs.2013.12.001.

[10] A. Prabha, J. Yadav, A. Rani, and V. Singh, "Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier," *Comput. Biol. Med.*, vol. 136, no. March, p. 104664, 2021, doi: 10.1016/j.compbiomed.2021.104664.

[11] Z. Zhang *et al.*, "A novel evolutionary ensemble prediction model using harmony search and stacking for diabetes diagnosis," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 1, p. 101873, 2024, doi: 10.1016/j.jksuci.2023.101873.

[12] H. Zhou, Y. Xin, and S. Li, "A diabetes prediction model based on Boruta feature selection and ensemble learning," *BMC Bioinformatics*, vol. 24, no. 1, pp. 1–34, 2023, doi: 10.1186/s12859-023-05300-5.

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

[14] Y. Su, C. Huang, W. Zhu, X. Lyu, and F. Ji, "Multi-party Diabetes Mellitus risk prediction based on secure federated learning," *Biomed. Signal Process. Control*, vol. 85, no. August, 2023, doi: 10.1016/j.bspc.2023.104881.

[15] P. Yadav, S. C. Sharma, R. Mahadeva, and S. P. Patole, "Exploring Hyper-parameters and Feature Selection for Predicting Non-communicable Chronic Disease using Stacking Classifier," *IEEE Access*, vol. 11, no. July, pp. 80030–80055, 2023, doi: 10.1109/ACCESS.2023.3299332.

[16] A. Doğru, S. Buyrukoğlu, and M. Arı, "A hybrid super ensemble learning model for the early-stage prediction of diabetes risk," *Med. Biol. Eng. Comput.*, vol. 61, no. 3, pp. 785–797, 2023, doi: 10.1007/s11517-022-02749-z.

[17] H. Hairani and D. Priyanto, "A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, pp. 585–590, 2023, doi: 10.14569/IJACSA.2023.0140864.

[18] D. Kumar, N. Mandal, and Y. Kumar, "Fog-based framework for diabetes prediction using hybrid ANFIS model in cloud environment," *Pers. Ubiquitous Comput.*, vol. 27, no. 3, pp. 909–916, 2023, doi: 10.1007/s00779-022-01678-w.

[19] C. N. Aher and A. K. Jena, "Improved invasive weed bird swarm optimization algorithm (IWBSOA) enabled hybrid deep learning classifier for diabetic prediction," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 4, pp. 3929–3945, 2023, doi: 10.1007/s12652-022-04462-z.

[20] A. Rajagopal, S. Jha, R. Alagarsamy, S. G. Quek, and G. Selvachandran, "A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures," *Math. Comput. Simul.*, vol. 198, pp. 388–406, 2022, doi: 10.1016/j.matcom.2022.03.003.

[21] B. S. Ahamed, M. S. Arya, and A. O. Nancy V, "Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques," *Front. Comput. Sci.*, vol. 4, no. May, pp. 1–5, 2022, doi: 10.3389/fcomp.2022.835242.

[22] S. BUYRUKOĞLU and A. AKBAŞ, "Machine Learning based Early Prediction of Type 2 Diabetes: A New Hybrid Feature Selection Approach using Correlation Matrix with Heatmap and SFS," *Balk. J. Electr. Comput. Eng.*, vol. 10, no. 2, pp. 110–117, 2022, doi: 10.17694/bajece.973129.

[23] A. A. Taha and S. J. Malebary, "A Hybrid Meta-Classifier of Fuzzy Clustering and Logistic Regression for Diabetes Prediction," *Comput. Mater. Contin.*, vol. 71, no. 2, pp. 6089–6105, 2022, doi: 10.32604/cmc.2022.023848.

[24] R. Ghabousian, Y. Farhang, K. Majidzadeh, and A. B. Sangarh, "Hybrid of particle swarm optimization algorithm and fuzzy system for diabetes diagnosis," *Int. J. Nonlinear Anal. Appl. Press*, vol. 6822, no. July, pp. 2008–6822, 2022, [Online]. Available: http://dx.doi.org/10.22075/ijnaa.2022.29575.4196.

[25] G. Mutlu, "SVM-SMO-SGD : A hybrid-SVM-SMO-SGD: A hybrid-parallel support vector machine algorithm using sequential minimal optimization with stochastic gradient descent," *Parallel Comput.*, vol. 113, no. July, 2022, doi: 10.1016/j.parco.2022.102955.

[26] M. O. Edeh *et al.*, "A Classification Algorithm-Based Hybrid Diabetes Prediction Model," *Front. Public Heal.*, vol. 10, no. March, pp. 1–7, 2022, doi: 10.3389/fpubh.2022.829519.

[27] X. Li, J. Zhang, and F. Safara, "Improving the Accuracy of Diabetes Diagnosis Applications through a Hybrid Feature Selection Algorithm," *Neural Process. Lett.*, vol. 55, no. 1, pp. 153–169, 2021, doi: 10.1007/s11063-021-10491-0.

[28] S. K. Kalagotla, S V Gangashetty, and K. Giridhar, "A novel stacking technique for prediction of diabetes," *Comput. Biol. Med.*, vol. 135, no. June, p. 104554, 2021, doi: 10.1016/j.compbiomed.2021.104554.

[29] N. Kanimozhi and G. Singaravel, "Hybrid artificial fish particle swarm optimizer and kernel extreme learning machine for type-II diabetes predictive model," *Med. Biol. Eng. Comput.*, vol. 59, no. 4, pp. 841–867, 2021, doi: 10.1007/s11517-021-02333-x.

[30] M. Dinesh and D.Prabha, "Diabetes Mellitus Prediction System Using Hybrid KPCA-GA-SVM Feature Selection Techniques," *J. Phys. Conf. Ser.*, vol. 1767, no. 1, p. 012001, 2021, doi: 10.1088/1742-6596/1767/1/012001.

[31] C. Mallika and S. Selvamuthukumaran, "A Hybrid Crow Search and Grey Wolf Optimization Technique for Enhanced Medical Data Classification in Diabetes Diagnosis System," *Int. J. Comput. Intell. Syst.*, vol. 14, no. 1, 2021, doi: 10.1007/s44196-021-00013-0.

[32] S. Samreen, "Memory-efficient, accurate and early diagnosis of diabetes through a machine learning pipeline employing crow search-based feature engineering and a stacking ensemble," *IEEE Access*, vol. 9, pp. 134335–134354, 2021, doi: 10.1109/ACCESS.2021.3116383.

[33] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *Int. J. Cogn. Comput. Eng.*, vol. 2, no. November 2020, pp. 40–46, 2021, doi: 10.1016/j.ijcce.2021.01.001.

[34] P. Rajendra and S. Latifi, "Prediction of diabetes using logistic

regression and ensemble techniques," *Comput. Methods Programs Biomed. Updat.*, vol. 1, p. 100032, 2021, doi: 10.1016/j.cmpbup.2021.100032.

[35] A. Singh, A. Dhillon, N. Kumar, M. S. Hossain, G. Muhammad, and M. Kumar, "eDiaPredict: An Ensemble-based Framework for Diabetes Prediction," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 17, no. 2s, 2021, doi: 10.1145/3415155.

[36] M. S. Islam, M. K. Qaraqe, S. B. Belhaouari, and M. A. Abdul-Ghani, "Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes," *IEEE Access*, vol. 8, pp. 120537–120547, 2020, doi: 10.1109/ACCESS.2020.3005540.

[37] S. Rajendar, R. Thangaraj, J. Palanisamy, and V. K. Kaliappan, "Comparative analysis of classifier models for the early prediction of type 2 diabetes," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 7, pp. 2184–2194, 2020.

[38] G. Tripathi and R. Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning," in *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (2020) 1009-1014*, 2020, pp. 1009–1014, doi: 10.1109/ICRITO48877.2020.9197832.

[39] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimed. Tools Appl.*, vol. 80, no. 5, pp. 8091–8126, 2021, doi: 10.1007/s11042-020-10139-6.

[40] P. Melillo, N. De Luca, M. Bracale, and L. Pecchia, "Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability," *IEEE J. Biomed. Heal. Informatics*, vol. 17, no. 3, pp. 727–733, 2013, doi: 10.1109/JBHI.2013.2244902.

[41] M. A. Peabody, T. Van Rossum, R. Lo, and F. S. L. Brinkman, "Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities," *BMC Bioinformatics*, vol. 16, no. 1, 2015, doi: 10.1186/s12859-015-0788-5.

[42] F. Beekmann, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," *Stichprobenbasierte Assoz. im Rahmen des Knowl. Discov. Databases*, pp. 5–50, 2003, doi: 10.1007/978-3-322-81227-8_2.

[43] D. C. Yadav and S. Pal, "Prediction of thyroid disease using decision tree ensemble method," *Human-Intelligent Syst. Integr.*, vol. 2, no. 1–4, pp. 89–95, 2020, doi: 10.1007/s42454-020-00006-y.

[44] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," *Proc. 6th Int. Conf. Inven. Comput. Technol. ICICT 2021*, pp. 1329–1333, 2021, doi: 10.1109/ICICT50816.2021.9358597.

[45] S. Meesri, S. Phimoltares, and A. Mahaweerawat, "Diagnosis of Heart Disease Using a Mixed Classifier," *ICSEC 2017 - 21st Int. Comput. Sci. Eng. Conf. 2017, Proceeding*, vol. 6, pp. 118–123, 2018, doi: 10.1109/ICSEC.2017.8443940.

[46] "Pima Indians Diabetes Database | Kaggle." https://www.kaggle.com/uciml/pima-indians-diabetes-database (accessed Jul. 29, 2021).

[47] G. Geetha and K. M. Prasad, "An Hybrid Ensemble Machine Learning Approach to Predict Type 2 Diabetes Mellitus," *Webology*, vol. 18, no. SpecialIssue2, pp. 311–331, 2021, doi: 10.14704/WEB/V18SI02/WEB18074.

[48] DaSilva John, "diabetes | Kaggle," 2022, Accessed: Apr. 23, 2022. [Online]. Available: https://www.kaggle.com/datasets/johndasilva/diabetes?resource=download.

[49] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.

[50] H. Naz and S. Ahuja, "SMOTE - SMO - based expert system for type II diabetes detection using PIMA dataset," *Int. J. Diabetes Dev. Ctries.*, vol. 42, no. June, pp. 245–253, 2022, doi: 10.1007/s13410-021-00969-x.

[51] D. K. Yadav, C. Azad, K. Bala, P. K. Sharma, and S. Kumar, "Genetic Algorithm and Naïve Bayes-Based (GANB) Diabetes Mellitus Prediction System," in *Lecture Notes in Electrical Engineering*, vol. 887, Springer, Singapore, 2023, pp. 561–572.

[52] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimed. Syst.*, no. 0123456789, 2021, doi: 10.1007/s00530-021-00817-2.

[53] P. Madan *et al.*, "An Optimization-Based Diabetes Prediction Model Using CNN and Bi-Directional LSTM in Real-Time Environment," *Appl. Sci.*, vol. 12, no. 8, 2022, doi: 10.3390/app12083989.

[54] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, "Diabetes detection using deep learning techniques with oversampling and feature augmentation," *Comput. Methods Programs Biomed.*, vol. 202, no. May 2022, 2021, doi: 10.1016/j.cmpb.2021.105968.

[55] Y. Wu *et al.*, "Novel binary logistic regression model based on feature transformation of XGBoost for type 2 Diabetes Mellitus prediction in healthcare systems," *Futur. Gener. Comput. Syst.*, vol. 129, Nov. 2021, doi: 10.1016/j.future.2021.11.003.

[56] R. M.S and M. Lakshmi, "Autonomous prediction of Type 2 Diabetes with high impact of glucose level," *Comput. Electr. Eng.*, vol. 101, p. 108082, Jul. 2022, doi: 10.1016/j.compeleceng.2022.108082.

[57] H. Qi, X. Song, S. Liu, Y. Zhang, and K. K. L. Wong, "KFPredict: An ensemble learning prediction framework for diabetes based on fusion of key features," *Comput. Methods Programs Biomed.*, vol. 231, p. 107378, Apr. 2023, doi: 10.1016/j.cmpb.2023.107378.

TABLE II
DIFFERENT PARAMETER VALUES (70-30 SPLIT)

	Accuracy	Recall	Precision	F-1	MCC	Kappa Value
SVM	0.796	0.802	0.655	0.721	0.571	0.564
LR	0.779	0.723	0.647	0.683	0.516	0.514
DT	0.974	0.921	0.960	0.958	0.941	0.939
NB	0.783	0.710	0.658	0.683	0.520	0.519
KNN	0.809	0.815	0.673	0.738	0.597	0.590
SMOTE, SMO [50]	0.990	0.982	0.962	0.977	-------	-------
GA, SMOTE, NB [51]	0.829	-------	-------	-------	-------	-------
SMOTE, GA, DT [52]	0.821	0.859	0.807	-------	-------	-------
KPCA, GA, SVM [30]	0.973	0.914	0.924	0.919	-------	0.890
PIMA(proposed model)	0.991	0.986	0.986	0.986	0.980	0.980

TABLE III
DIFFERENT PARAMETER VALUES WITH TEN-FOLD CROSS VALIDATION

	Accuracy	Recall	Precision	F-1	MCC	Kappa Value
SVM	0.778	0.822	0.757	0.787	0.560	0.556
LR	0.737	0.708	0.752	0.729	0.475	0.474
DT	0.764	0.794	0.749	0.770	0.529s	0.528
NB	0.741	0.692	0.768	0.727	0.484	0.482
KNN	0.779	0.856	0.742	0.794	0.566	0.558
Boruta, K-means, NB, KNN, DT, SVM [12]	0.981	0.984	0.977	0.980	0.965	0.962
SMOTE, Boruta, Grid Search, Grey Wolf[15]	0.963	0.971	0.982	0.980	0.743	-------
CNN-Bi-LSTM [53]	0.988	0.940	0.980	0.960	-------	0.940
VAE + SAE With CNN [54]	0.932	-------	-------	-------	-------	-------
X-BLR [55]	0.940	0.940	0.920	0.930	-------	-------
CGLSTM [56]	0.978	0.896	0.914	0.856	-------	-------
KF Predict [57]	0.935	0.980	0.850	-------	-------	-------
PIMA(proposed model)	0.990	0.987	0.988	0.987	0.987	0.980

TABLE IV
DIFFERENT PARAMETER VALUES (70-30 SPLIT) FOR RAW DATA

	Accuracy	Recall	Precision	F-1	MCC	Kappa Value
SVM	0.735	0.487	0.661	0.561	0.387	0.378
LR	0.740	0.625	0.625	0.625	0.426	0.426
DT	0.727	0.650	0.597	0.622	0.410	0.409
NB	0.744	0.662	0.623	0.642	0.444	0.444
KNN	0.688	0.562	0.548	0.555	0.315	0.315
PIMA (proposed model)	0.774	0.550	0.733	0.628	0.481	0.471

TABLE V
DIFFERENT PARAMETER VALUES WITH TEN-FOLD CROSS-VALIDATION FOR RAW DATASET

	Accuracy	Recall	Precision	F-1	MCC	Kappa Value
SVM	0.727	0.457	0.652	0.562	0.435	0.411
LR	0.735	0.529	0.621	0.602	0.480	0.422
DT	0.720	0.526	0.599	0.597	0.388	0.385
NB	0.747	0.574	0.688	0.608	0.453	0.425
KNN	0.720	0.553	0.618	0.580	0.376	0.372
PIMA (proposed model)	0.753	0.550	0.706	0.609	0.609	0.433

TABLE VI
STANDARD DEVIATION VALUES FOR DIFFERENT CLASSIFIERS

	Accuracy	Recall	Precision	F-1	MCC	Kappa Value
SVM	0.04618	0.11605	0.07477	0.10878	0.11951	0.12548
LR	0.06523	0.12613	0.10740	0.11924	0.15650	0.15818
DT	0.05074	0.12894	0.06717	0.09107	0.12096	0.12122
NB	0.05669	0.12838	0.08984	0.10335	0.13006	0.13449
KNN	0.04267	0.08270	0.07377	0.06319	0.09176	0.09121
PIMA (proposed model)	0.07987	0.11456	0.18476	0.11985	0.11985	0.17749

TABLE VII
DIFFERENT PARAMETER VALUES (70-30 SPLIT)

	Accuracy	Recall	Precision	F-1	MCC	Kappa Value
SVM	0.805	0.826	0.678	0.745	0.597	0.589
LR	0.743	0.734	0.605	0.663	0.464	0.459
DT	0.985	0.992	0.975	0.982	0.988	0.988
NB	0.775	0.719	0.659	0.688	0.513	0.512
KNN	0.950	0.975	0.876	0.932	0.897	0.892
Frankfurt (proposed model)	0.998	0.998	0.993	0.995	0.994	0.996

TABLE VIII
DIFFERENT PARAMETER VALUES WITH TEN-FOLD CROSS-VALIDATION

	Accuracy	Recall	Precision	F-1	MCC	Kappa Value
SVM	0.800	0.827	0.785	0.805	0.602	0.604
LR	0.734	0.705	0.750	0.726	0.470	0.468
DT	0.973	0.963	0.983	0.972	0.947	0.946
NB	0.731	0.679	0.759	0.716	0.465	0.462
KNN	0.878	0.967	0.823	0.888	0.770	0.757
Frankfurt (proposed model)	0.995	0.988	0.990	0.990	0.990	0.990

TABLE IX
DIFFERENT PARAMETER VALUES (70-30 SPLIT) FOR RAW DATA

	Accuracy	Recall	Precision	F-1	MCC	Kappa Value
SVM	0s.788	0.559	0.776	0.650	0.518	0.504
LR	0.805	0.625	0.776	0.692	0.559	0.552
DT	0.961	0.981	0.915	0.947	0.918	0.917
NB	0.795	0.663	0.729	0.694	0.542	0.540
KNN	0.813	0.690	0.748	0.718	0.580	0.579
Frankfurt (proposed model)	0.961	0.981	0.915	0.947	0.918	0.917

TABLE X
DIFFERENT PARAMETER VALUES WITH TEN FOLD CROSS-VALIDATION FOR ORIGINAL DATA

	Accuracy	Recall	Precision	F-1	MCC	Kappa Value
SVM	0.764	0.486	0.730	0.581	0.445	0.427
LR	0.761	0.522	0.700	0.595	0.443	0.432
DT	0.945	0.921	0.918	0.918	0.878	0.877
NB	0.751	0.573	0.653	0.608	0.431	0.428
KNN	0.782	0.654	0.695	0.672	0.512	0.510
Frankfurt (proposed model)	0.961	0.981	0.918	0.948	0.948	0.917

TABLE XI
STANDARD DEVIATION VALUES FOR DIFFERENT CLASSIFIERS

	Accuracy	Recall	Precision	F-1	MCC	Kappa Value
SVM	0.02857	0.05319	0.07273	0.05086	0.07127	0.06811
LR	0.03443	0.06605	0.08155	0.05951	0.08455	0.08144
DT	0.02214	0.04647	0.04193	0.03218	0.04852	0.04896
NB	0.03270	0.06243	0.06337	0.05221	0.07354	0.07347
KNN	0.02842	0.06161	0.04705	0.04413	0.06504	0.06433
Frankfurt (proposed model)	0.02242	0.02307	0.04708	0.02960	0.02960	0.04750

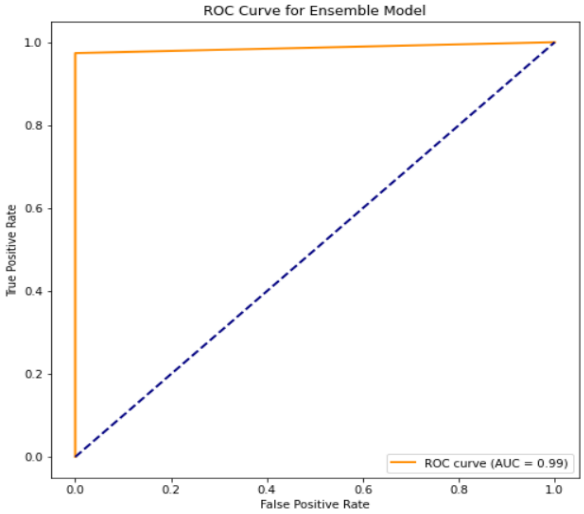

Fig. 2. ROC Curve for PIMA Indian Diabetes Dataset

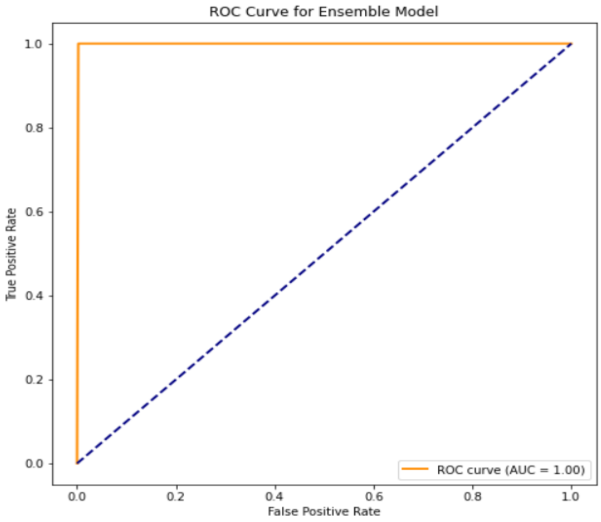

Fig. 3. ROC Curve for Hospital Frankfurt Diabetes Dataset

TABLE XII

DIFFERENT ERROR RATES FORMULAS

Error Rates	Formulas
RSE	$\frac{\text{Sum of Squared Errors of Model}}{\text{Sum of Squared Errors of Baseline Model}}$
RRSE	$\sqrt{RSE}$
MAE	$\frac{\text{Sum of Absolute Errors}}{\text{Number of Data Points}}$
RMSE	$\sqrt{\text{Mean of Squared Errors}}$
MSTSS	$\frac{\text{Sum of Squared Differences from Mean}}{\text{Number of Data Points}}$

TABLE XIII

VALUES OF ERROR RATES FOR BOTH DATASETS

Error Rates for the Proposed Method	PIMA	Hospital Frankfurt
RSE	0.019	0.036
RRSE	0.140	0.192
MAE	0.008	0.0016
RMES	0.093	0.040
MSTSS	0.22	0.226