IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 June 2023, accepted 21 July 2023, date of publication 27 July 2023, date of current version 3 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3299332

RESEARCH ARTICLE

Exploring Hyper-Parameters and Feature Selection for Predicting Non-Communicable Chronic Disease Using Stacking Classifier

POOJA YADAV[1,2], (Member, IEEE), S. C. SHARMA[1], RAJESH MAHADEVA[3,4], (Member, IEEE), AND SHASHIKANT P. PATOLE[3], (Member, IEEE)

[1]Electronics and Computer Discipline, DPT, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand 247667, India
[2]Department of Computer Science and Information Technology, FET, M. J. P. Rohilkhand University, Bareilly, Uttar Pradesh 243006, India
[3]Department of Physics, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
[4]Division of Research and Innovation, Uttaranchal University, Dehradun 248012, India

Corresponding authors: S. C. Sharma (subhash.sharma@pt.iitr.ac.in), Shashikant P. Patole (shashikant.patole@ku.ac.ae), and Rajesh Mahadeva (rajeshmahadeva15@gmail.com)

This work was supported by the Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates, under the Khalifa University Research Internal Fund.

ABSTRACT Non-communicable disease, especially chronic disease, is the most common factor of complication of deteriorating physical health and the state of one's mind. It is also a prominent cause of illness and mortality around the world. Primarily chronic disease is preventable at a particular stage though its occurrence is critical. To make clinical decisions, these illness prediction models were created to assist clinicians and patients. A chronic disease prediction model takes into account many risk variables to determine an individual's illness risk. Machine learning approaches have made it possible to predict chronic disease early by collecting Electronic Health Record (EHR) data. This paper focuses on the diabetes dataset extracted from Kaggle and two unseen real datasets. In this paper, we have implemented Synthetic Minority Over-Sampling Technique (SMOTE) algorithm to balance the dataset. We have also explored Boruta as the feature selection method. To tune hyper-parameters of different algorithms, we have proposed an improved technique by combining the Grid Search method with the Grey Wolf Optimization algorithm. The Grid Search method requires extensive searching, while the Grey Wolf Optimization algorithm is easily linked, rapid to seek, and extremely exact. Nine conventional classification techniques have been evaluated in this paper. This research concentrates on the Stacking Classifier to assess the performance of the prediction model that produces the best results. The Proposed Model gave the highest F1-Score 98.84% on PIMA dataset, 98% after validation on the Synthetic dataset, 97.3% on ADRC dataset, 96.20% on FHD dataset. To the best of our knowledge, no previous work has focused on such a sort of technique and these two datasets. The outcomes of the comparison experiment on the PIMA dataset reveals that the proposed strategy performs better. This study also provides the interpretation of the proposed model. It conducts an ethical assessment of what explainability means for the use of Machine Learning models in clinical practice.

INDEX TERMS Chronic disease, feature selection, hyperparameter tuning, machine learning, non-communicable diseases, prediction model, stacking classifier, interpretability.

I. INTRODUCTION

Machine Learning (ML) is ''a field of study that enables computers to learn without being explicitly programmed.'' An IBM employee who was also an American pioneer coined, defined, and emerged this effective term 'machine learning' in 1959. However, with broad implementation, ML has advanced dramatically in the previous decade. Artificial Intelligence (AI), Data Mining & ML were early adopters in health care services by building robots that can analyze, learn,

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram.

and respond to data, and their popularity has since expanded with the advent of genetic data and, most importantly, the broad availability of wearable sensors. Deep Learning (DL) is a very recent topic in ML [1]. The integration of health sciences, computer science, and information science to manage and disseminate data for clinical practice is known as clinical informatics. Furthermore, clinical informatics tools are resources that facilitate health practitioners to quickly gather, share, and use data and ICT knowledge to improve healthcare delivery. The tools make it easier to integrate ICT to help patients and doctors make decisions while promoting evidence-based medicine. Daily, using traditional methodologies, it is extremely difficult to examine and manage massive amounts of diverse data and information generated by healthcare providers. ML/DL approaches assist in properly analyzing this data for meaningful insights. Furthermore, a variety of data sources, including genomes, health data, social media data, and climate data, might be used to improve healthcare data. The four major sectors of healthcare that can be improved by ML / DL approaches are prognosis, diagnosis, medication, and clinical workflow [2].

Most databases contain personally identifiable information, including health information, that's why these are inaccessible. Identifiable documents are complicated to share since organizations must adhere to strict guidelines. When it comes to acquiring critical datasets, researchers and analysts continue to confront several challenges. Data access criteria include the requirement for data usage agreements, the preparation and authorization of a comprehensive protocol, the finalization of a data request form, the acceptance of an ethical assessment, and the cost of getting datasets that are not in the public domain, remain a challenge [3]. So, Synthetic data, or artificially created data, has recently been fascinated because of its potential for making timely healthcare data more available for research and technology development [3]. Prediction is nothing new in healthcare since clinical practitioners are adept at assessing risk factors or creating data-driven prognostic forecasts. Machine learning techniques can outperform classic regression models in terms of prediction accuracy [4]. A prediction model is a model that provides a way to assess a patient's risk of disease outcome. With the proliferation of such prediction models, the question of when, what and how to use them arises. Depending on the company's needs, these models can be tried to teach over time to respond to new information or perspectives [5]. When patients fail to define their medical problems correctly, based on laboratory research, it can result in some probability of error. Healthcare professionals struggle to make choices about illnesses because they may lack expert knowledge in some areas. To overcome this limitation, it is necessary to create a disease prediction system by combining knowledge of medicine with an integrated approach to produce the best results and benefit society [6]. Diabetes is classified as a non-communicable disease (NCD),i.e. it does not spread from one person to another. NCDs are medical conditions that last for a long time and progress slowly. Genetic, physiological, behavioral, lifestyle, and environmental factors are the leading cause of NCDs. According to the World Health Organization (WHO), NCDs are the leading cause of death worldwide, accounting for 71% of all deaths yearly. Illness self-awareness in patients is crucial for disease control yet challenging to attain because NCDs are chronic, hidden, and irreversible. Cardiovascular diseases (CVD), cancers, respiratory diseases, liver, and chronic kidney disease (CKD) are the top killers among NCDs [7], [8]. Diabetes causes other health problems. High and low blood pressure, nerve damage, and bone difficulties are all symptoms of CVD and CKD. In which, Diabetes, high blood pressure, and CVD are all risk factors for CKD patients, according to [9]. AI has become a viable method for creating computer-aided diagnostics (CAD) in the field of medicine [10], [11], which could be used to explore hidden associations between the onset of CKD and the onset of its symptoms, allowing for the early identification of patients at risk. Diabetes Mellitus (DM) is a long-term state in which the body fails to generate or even use insulin properly. It is caused by genetic predisposition, poor lifestyle, and unhealthy diet. Diabetes is a well-known risk factor for CVD. Individuals who have type 2 DM(T2DM) have a higher risk of CVD mortality and morbidity than those who are not diabetic [12], [13].

In this study, we worked on one public dataset i.e. PIMA and Two real Indian datasets from S.N. Medical College, Agra, Uttar Pradesh, India. It has 583 samples. And another is the 'FHD dataset' collected from Future Hospital, Bareilly, Uttar Pradesh, India. It has 400 samples.

A. CONTRIBUTION

This research aims to develop an effective method to predict chronic diseases, like diabetes, as accurately as possible. The following are the paper's main contributions:

- The PIMA [14] data set is used, which is unbalanced. We validated the proposed model using a synthetic dataset of 500 samples based on PIMA. We have also worked on two Indian and actual diabetes mellitus datasets. These private and real datasets are referred to as the 'ADRC dataset' and 'FHD dataset' in this work. ADRC dataset was received from Professor (Dr.) Prabhat Agrawal S.N. Medical College Agra, while FHD was collected from Future Hospital Bareilly.
- We have applied a balancing algorithm: Synthetic Minority Over-sampling TechniquE (SMOTE), to balance the unbalanced dataset.
- Evaluate the effectiveness of the various modeling techniques using a set of full features along with a set of extracted features after exploring approaches such as Boruta.
- Optimize hyperparameters using Grid Search and also apply the proposed Hybrid approach, GS-GWO, to optimize hyperparameters and find the best hyperparameters.

TABLE 1. Acronyms and their meanings.

Acronyms	Explanation	Acronyms	Explanation
AB	Adaboost	JCC	Joint Clustering And Classification
AI	Artificial Intelligence	KNN	K-Nearest Neighbour
APDFS	Adaptive Probabilistic Divergence-Based Feature Selection	LR	Logistic Regression
AUC	Area Under Curve	LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory	MCC	Matthews Correlation Coefficient
CAD	Computer-Aided Diagnosis	MICE	Multiple Imputations By Chained Equations
CART	Classification And Regression Trees	ML	Machine Learning
CKD	Chronic Kidney Disease	MLP	Multilayer Perceptron
CNN	Convolutional Neural Network	MZSA	Maximum Z Score Among Shadow Attributes
CSV	Comma Separated Value	NB	Naive Bayes
CVD	Cardiovascular Diseases	NCD	Non-Communicable Diseases
DCNN	Deep Convolutional Neural Network	NLPNN	Network-Limited Polynomial Neural Network
DL	Deep Learning	NN	Neural Network
DM	Diabetes Mellitus	PCA	Principal Component Analysis
DT	Decision Tree	PNN	Probabilistic Neural Networks
EDA	Exploratory Data Analysis	PPV	Positive Predictive Value
EHR	Electronic Health Record	RF	Random Forest
FN	False Negative	RMSE	Root Mean Square Error
FOS	First-Order Statistics	ROC	Receiver Operating Characteristic
FP	False Positive	SMOTE	Synthetic Minority Over-Sampling Technique
FPR	False Positive Rate	SVM	Support Vector Machines
GB	Gradient Boosting	T2DM	Type 2 DM
GLCM	Gray-Level Co-Occurrence Matrix	TN	True Negative
GNN	Graph Neural Networks	TP	True Positive
GS-GWO	Grid Search- Grey Wolf Optimization	TPR	True Positive Rate
GWO	Grey Wolf Optimization	UCI	University Of California, Irvine
HLRM	Hyper-Parameterized Logistic Regression Model	WHO	World Health Organization
ID3	Iterative Dichotomiser 3	WPN	Weighted Patient Network
IQR	Inter Quartile Range	XB	Xgboost

- We have proposed an efficient prediction model based on a stacking classifier to enhance the outcome.
- To comprehend how the model predicts the decision, explainable AI algorithms such as LIME and SHAP are applied. These methods aid in determining which traits are most important in prediction.

B. STRUCTURE OF THE PAPER

The remainder of the article is organized as follows: Table 1 contains a list of all the acronyms used in this work. Section II is a literature review that discusses past work. Section III defines methodology which contains a brief idea of all the datasets, feature scaling, balancing algorithm: SMOTE, feature selection method, hyper-parameter tuning, classification algorithms, and proposed model. Experimental environment and results in terms of performance evaluation, discussion and statistical test are presented in Section IV in detail. Section V discusses the interpretation of the proposed model. Also, comparisons with previous work have been focused on Section VI. Section VII elaborates discussion part and Section VIII makes some concluding remarks.

II. LITERATURE REVIEW

This section discusses the various AI and ML techniques used recently to detect and diagnose chronic diseases. It is very hard to predict which features will be most significant. Data is often acquired by describing occurrences with as many details as possible and then deciding which are significant. Too many features might be harmful as a result of which, the effect of significant differences and the decision model's similarities will diminish as it attempts to integrate all possible information. Techniques of feature selection seek the minimal number of features that yield the best classification [15]. Many features can overburden the classifier, significantly impacting classification calculation and lengthening the computational time. Many features might overwhelm the classifier, affecting classification calculation and increasing the calculation time. The goal of choosing feature subsets is to reduce calculation time while improving prediction outcomes by deleting features/attributes from a dataset deemed uninteresting or incapable of contributing to classification accuracy [16]. Authors [17] proposed a distinct feature selection scheme coupled with a machine-learning technique that may quickly detect a premature chronic illness. As a direct consequence, a new strategy adaptive probabilistic divergence-based feature selection (APDFS) strategy for the earlier detection of chronic disease, is offered in conjunction with the hyper-parameterized logistic regression model (HLRM). The APDFS algorithm identifies characteristics that are important for CKD diagnosis. The data set containing just the specified attributes is fed to the HLRM model, which is utilized to anticipate chronic illness in its early stages with 91.6% accuracy. Authors [18] proposed BSWE GWO KELM, a feature selection framework for predicting intra-dialectic hypotension utilizing chronic kidney disease-mineral and bone problems.

Authors [19] proposed Gray-Level Co-Occurrence Matrix (GLCM) and First-Order Statistics (FOS) for feature selec-

tion with a voting classifier to predict fatty liver and achieved 97.1% accuracy. The authors [20] developed a prediction model that employed logistic regression to predict the probability of the emergence of liver disorders. The predictive model performed admirably, with a forecast accuracy rate of 72.4%. Through prediction, the study [21] evaluates the effectiveness of various ML algorithms in reducing the high cost of detection of chronic liver disease. With a 75% accuracy rate, logistic regression produces the best results. This work's [22] demonstrated a novel knowledge-based system based on fuzzy rules generated by Classification and Regression Trees (CART). The findings imply that integrating fuzzy rule-based CART with de-noising and clustering techniques increases sickness prediction accuracy and efficacy from real-world medical datasets. This research [23] is focused on identifying essential predictors of liver disease. To generate missing data points, multiple imputations by chained equations (MICEs) were employed, while for dimensionality reduction, principal component analysis (PCA) was employed. Among the various algorithms, Random Forest provided the highest accuracy score of 98.14%. The authors [24] concentrated on developing a prediction model to detect risk variables for diabetes illness. The total accuracy of the machine learning-based system is 90.62%. For the K-10 fold method, the RF-based classifier with feature selection method based on LR provides 94.25% accuracy and 0.95 Area Under Curve (AUC).

The fundamental purpose of this study [25] is to create a system capable of reliably predicting diabetes in patients. The experimental findings show that the targeted framework with Ensemble Voting Classifier can achieve an accuracy of almost 86%. To adequately retrieve high-level signals concealed in chronic illness datasets, the authors [26] present a network-limited polynomial neural network (NLPNN) technique. This method augments data in feature space and aids in the reduction of over-fitting. The suggested approach may aid in the prompt and accurate diagnosis of chronic illnesses at an early stage. It can provide incredible precision results. The authors [27] propose a system based on Graph Neural Networks (GNNs) for forecasting chronic illness. To begin, researchers project a patient-disease bipartite graph used to create a Weighted Patient Network (WPN) that preserves the implicit link between patients. The prediction models are then built using GNN-based approaches. These models leverage WPN attributes to make strong patient representations for chronic illness prediction. Authors [28] suggest a classifier based on the ensemble method to improve decision-making for the identification of renal illness in an effective manner. Ensemble approaches integrate many learning algorithms to produce higher prediction performance than each constituent learning algorithm could accomplish. Furthermore, data is examined using 10-fold cross-validation, and system performance is measured using the receiver operating characteristic curve. Authors [29] tackled the prediction problem as a binary classification problem, and machine learning algorithms such as kernelized and sparse support vector machines (SVMs), Random forests, and sparse logistic regression were studied. They introduce two novel methods: K-LRT, a likelihood ratio test-based strategy, and JCC, a joint clustering and classification method that finds hidden patient groupings and tailors classifiers to each group. Reference [30] introduces a disease risk prediction technique by using a novel convolutional neural network (CNN)-based multimodal for structured and unstructured hospital data. When compared to other conventional prediction algorithms, 94.8% accuracy is achieved by the proposed method and a faster convergence time than the CNN-based unimodal illness risk prediction method.

Author [31] has applied different classifiers to predict the stage of chronic kidney disease. Among all Probabilistic Neural Networks (PNN) provide better accuracy. This research [32] uses a neural network (NN) model to investigate the challenge of predicting renal disease in hypertensive individuals. A hybrid neural network is also described, which combines Bidirectional Long Short-Term Memory (BiLSTM) and Autoencoder networks. The proposed approach [33] employs predictive analysis to identify the factors that fail in the early identification of Diabetic Mellitus. The decision tree algorithm and the Random forest have a maximum specificity of 98.20%, according to the data. Authors [34] worked on nine classification algorithms and compared all predictive models; among them, AUC for logistic regression was.8733, with sensitivity and specificity of.83 and.82, respectively. Using Magnetic Resonance Imaging (MRI), this study [35] provides a model prediction of LGG molecular subtypes. MR images were segmented and transformed into radiomics characteristics, providing predictive information on the classification of brain tumors.

A. RESEARCH GAP

In the literature, various concepts, methods, and techniques have been applied to different datasets, most of which belong to diseases and getting real medical data is another challenging task. A thorough study explains a high need to explore and select the best features as well as optimization techniques for hyperparameters. It has been observed that most of the medical datasets are unbalanced. Transforming an unbalanced class into a balance class is necessary to improve the functionality of the model of a classifier.

III. MATERIAL AND METHODS

The study technique and a dataset will be explained in this part. Working of the block diagram, Fig.1 are summaries in terms of the following points:

1) First of all PIMA dataset has been collected from Kaggle [14] public repositories and real datasets from Hospital.
2) Apply Exploratory Data Analysis and find out the null values, duplicate values, and outliers and handle them with the proper concept of data preprocessing.
3) Apply SMOTE to balance the dataset.

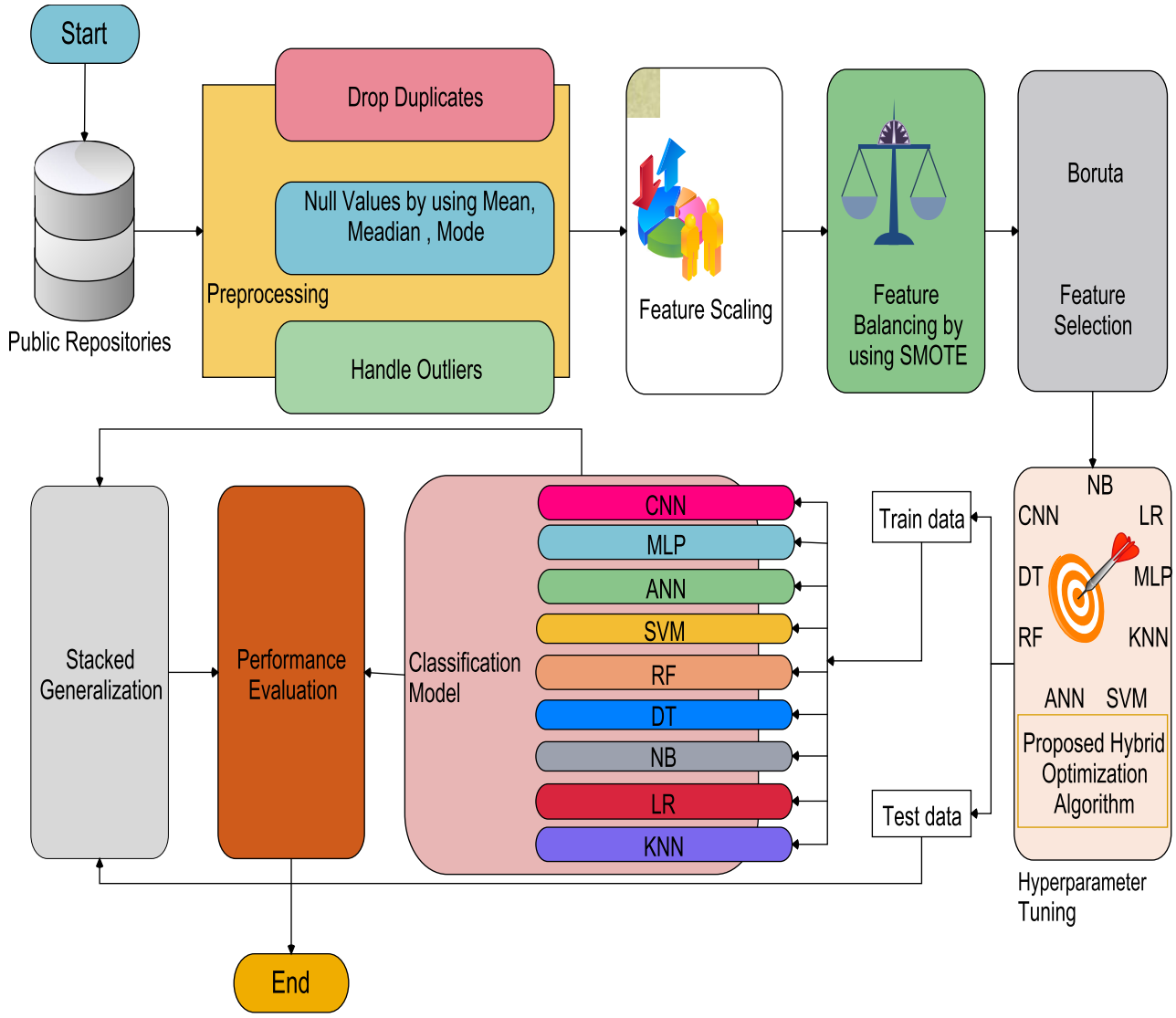


FIGURE 1. Work flowgraph.

4) Apply different Feature Selection and Extraction Methods.
5) Machine-learning models have been applied.
6) Apply training and testing.
7) Evaluate the performance.
8) Implement a Stacking classifier to improve the performance.
9) Analyze the performance of the classifier.
10) Interpret the Model.

A. DATASET

In this paper, we consider three diabetes disease datasets as shown in Table 2, one is a popular PIMA dataset [14], its synthetic dataset for validation, and two real diabetes disease datasets. These datasets belong to India. The dataset's description is as follows: This data set contains 768 patient data, 268 diabetic and 500 non-diabetic patients were found across nine columns. In these columns, eight are independent variables containing numerical data, and one is a dependent variable containing categorical value as mentioned in Table 3.

TABLE 2. Descriptions of all three diabetes dataset.

S.N.	Dataset	Data Types	Attributes	Instances	Having Disease	Not Having Disease
1	PIMA	Int, Float, Categorical	9	768	268	500
2	ADRC	Int, Float, Categorical	7	583	167	416
3	FHD	Int, Float, Categorical	7	400	150	250

TABLE 3. Attribute descriptions in the PIMA Dataset.

S.N.	Attribute	Description	Category	Scale
1	preg	Number of times pregnant	Numerical	
2	gluc/ GTT	Plasma glucose concentration at 2 Hours in an oral glucose tolerance test	Numerical	mg/dL
3	bP	Diastolic Blood Pressure	Numerical	mm Hg
4	sft	SkinTriceps skin fold thickness	Numerical	mm
5	insulin	Insulin2-Hour Serum insulin	Numerical	µh/mL
6	bmi	Body mass index	Numerical	Kg/m
7	dpf	Diabetes pedigree function	Numerical	
8	age	Age of the patient	Numerical	years
9	outcome	Binary value indicating diabetic/ non-diabetic Factor	Categorical	Yes ,No

TABLE 4. Attribute descriptions in the ADRC Dataset.

S.N.	Attribute	Description	Category	Scale
1	Age	Age of the Patient	Numerical	Years
2	Gender	Gender of Patient	Categorical	Male, Female
3	Neuropathy	Nerve Damage Symptom	Numerical 1-Yes, 0-No	1,0
4	Insulin	Body makes insulin	Numerical 1-Normal, 0-Not Normal	mIU/mL
5	Hb	Hemoglobin 12.0-15.0 for Female, 14-18 for Male	Numerical	gm/dl
6	HbA1c	Glycosylated Hemoglobin A1c Normal range is < 6.0	Numerical	%
7	Outcome	It indicates the result of the test	Categorical-Diabetic, Non Diabetic	Yes, No

Table 4 defines the attribute discription of ADRC dataset. It has seven attributes, of which six are independent and one is a dependent attribute. It contains 583 samples. The Neuropathy attribute indicates nerve damage symptoms occur if the patient has diabetes. Insulin is a hormone produced by our bodies to maintain appropriate blood glucose levels. It facilitates the entry of blood glucose (blood sugar) into your cells, which may be utilized for energy. This test indicates whether the insulin that our body produces is in the normal range or not. Hemoglobin tests are performed as part of a complete blood count (CBC), specifically red blood cells in the body that transports oxygen from the lungs to the rest of the body. While HbA1c is hemoglobin A1c test determines the average blood sugar level over the previous 2 to 3 months. It is also known as the glycated hemoglobin test and glycohemoglobin. Table tab:FHD represents the description of FHD dataset. It has 400 samples. Postprandial or postprandial glucose levels refer to blood sugar levels after eating. It should be less than 140. Two numbers are used to calculate blood pressure: The first number, systolic blood pressure, measures how much pressure is in your arteries when your heart beats. It should be less than 120. The second number, diastolic blood pressure, measures the pressure in your arteries between heartbeats. It should be less than 80.

TABLE 5. Attribute descriptions in the FHD Dataset.

S.N.	Attribute	Description	Category	Scale
1	Age	Age of the Patient	Numerical	Years
2	Gender	Gender of Patient	Categorical	Male, Female
3	Glucose	Post Prandial (PP) Blood Sugar Test. Normal range is < 140	Numerical	mg/dl
4	Systolic Blood Pressure	Normal range is <120	Numerical	mmHg
5	Diastolic Blood Pressure	Normal range is <80	Numerical	mmHg
6	Weight	Weight of the Patient	Numerical	Kg
7	Outcome	It indicates the result of the test	Categorical Diabetic, Non Diabetic	Yes , No

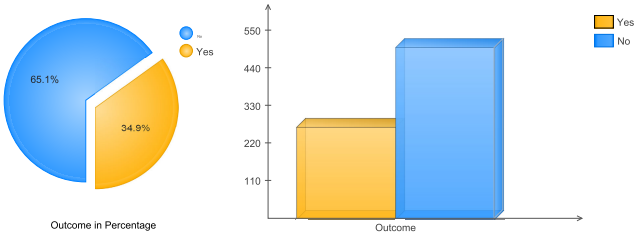
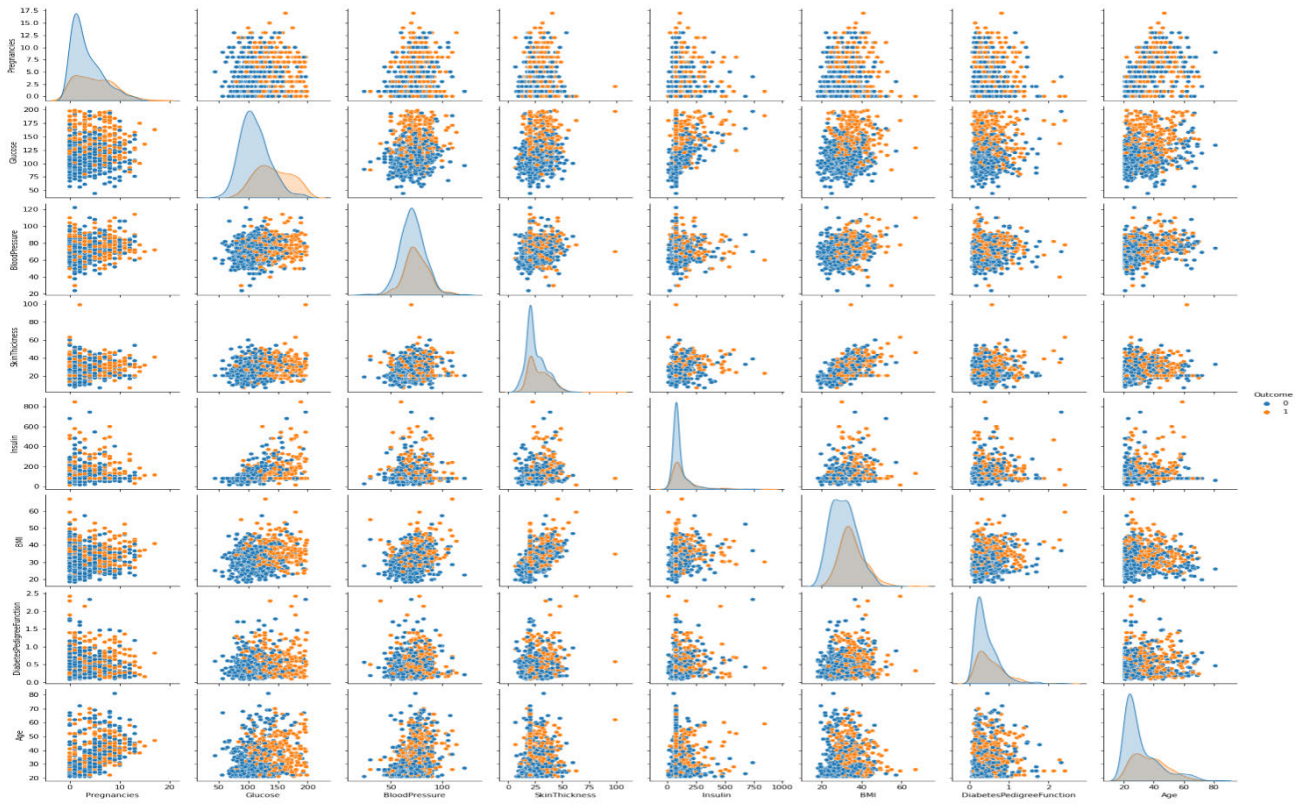


FIGURE 2. Countplot of outcome.

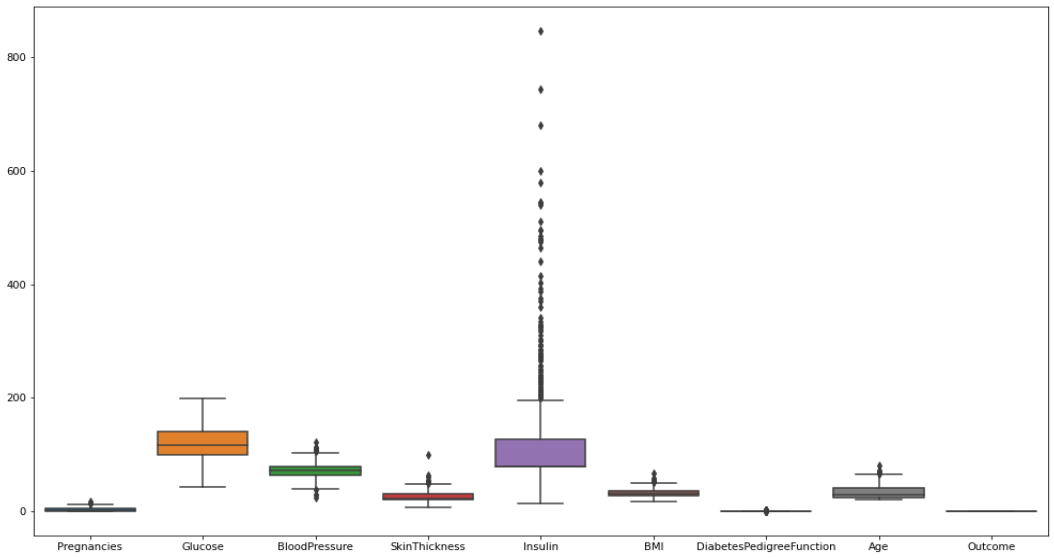
B. EXPLORATORY DATA ANALYSIS (EDA)

During EDA, it has been observed that:

- In our PIMA dataset in Fig.2, the proportion of diabetes patients is approximately half that of non-diabetic patients. This is a major imbalance. Similarly, in ADRC and FHD dataset, no. of diabetic patients are less than that of non-diabetic patients, so an imbalance exists.
- The original data format is CSV (Comma Separated Value).
- All three datasets do not have a null value, however, PIMA does have numerous zeros, such as:
  Glucose has a total of 5 zero values. Blood pressure has 35 zero values. The number of zero values in SkinThickness is 227, Insulin has 374 zero values, and BMI has 11 zero values, but again zero is a value and is handled by replacing it with the mean value.
  Similarly, ADRC has zero value in neuropathy, Insulin. While FHD has no zero value.
- Using the pair plot approach, Fig.3 illustrates the correlations between all features with respect to one another in seaborn. The color attribute could be utilized to make distinctions between each feature's various classifications. The 'Outcome' of the color used comes from the dataset.
- Outlier is the extreme value beyond the range of upper limit and lower limit. In PIMA, some outliers were detected with the help of Boxplot as mentioned in



(a) Relationship among all attributes



(b) Outliers

FIGURE 3. **Analysis of attributes.**

Fig.3b. These outliers could be the result of various underlying causes. It is advisable to normalize the data to protect against the negative consequences of the outliers. Due to the short size of the dataset, it was not preferred to eliminate unnecessary rows. They were handled by using IQR (Inter Quartile Range) method.

For assessing the original PIMA, a Synthetic dataset is created to validate our model. We compare variable distributions from 500 data samples generated using Gretel's Actgan Cloud based AI model to the original ground truth data under settings for dealing with missing data and outliers, including biased, imbalanced, small sampled data

and maintaining similarities. Fig.4 depicts the outcome of Gretel Report [36], [37].

C. FEATURE SCALING

This process is applied to scale a set of variables or features in data in an equal manner. In data processing, it is commonly referred to as data normalization or data standardization. Before utilizing machine learning methods to train models, feature scaling is often done at the last step of data preprocessing. It is intended to alter the data so that each characteristic falls inside a given range (e.g., between 0 and 1). As a result, training and tuning become more efficient and prevent any one characteristic from outperforming the others. There are several ways to scale features including, Normalisation (min-max scaling), Standardisation (z-score standardization), and decimal scaling (robust scaling). In this work, we have applied Standardisation to scale the features.

D. FEATURE BALANCING

The imbalance of class instances in the healthcare data set is a major problem. This indicates that the instance is not being divided appropriately across the various classes. Consequently, skewed class data classification results deliver a skewed conclusion in favor of the dominant class. To improve the performance of any machine learning algorithm on an unbalanced classification issue, data sampling or data balancing techniques adjust the ratio of class instances. There are two approaches to balancing uneven data collection, which is both under-sampling and over-sampling [38]. There are three fundamental methods for getting balanced data: (1) more samples from the minority group (oversampling); (2) fewer samples drawn from the majority group (undersampling); and (3) a combination of under and over sampling. Data oversampling entails duplicating instances of the minority class or synthesizing new minority class examples from existing ones. Oversampling methods for data include Random Oversampling SMOTE, Borderline SMOTE, SVM SMOTE, k-Means SMOTE, and ADASYN. The most used approaches are SMOTE and its variants, such as Borderline SMOTE. The quantity of oversampling to execute is maybe the most critical hyperparameter to tune. The undersampling strategy selects data at random or uses an algorithm to select which samples to remove from the majority class. However, this procedure results in data loss, which might impact the learning process [39]. Some Data undersampling methods are Random Undersampling, Condensed Nearest Neighbor, Tomek Links, Edited Nearest Neighbors, Neighborhood Cleaning Rule, One-Sided Selection. Modified nearest neighbors and Tomek linkages are two popular editing algorithms. All oversampling approaches can be used in pairing with almost every undersampling method. As a result, it may be advantageous to try various kinds of oversampling and undersampling strategies. A combination of some popular oversampling and undersampling approaches are: Random Undersampling and SMOTE, Tomek and SMOTE Links, SMOTE and Nearest Neighbours Edited. Depending on the machine learning method used, data sampling techniques may perform differently. Furthermore, the k-nearest neighbor approach is used internally by the majority of data sampling algorithms. The data types and sizes of input variables are extremely important to this method. As a result, it may be necessary to at least normalize input data with different scales before evaluating the approaches, and it may be necessary to use specialized methods if certain input variables are categorical rather than numerical. But in this paper, SMOTE with KNN algorithm has been implemented.

1) SMOTE

In this method, the minority of samples are now made up of synthetic ones. It raises the proportion of the minority class. Minority class to achieve equilibrium with the dominant class [40]. A ''balancer'' is just another term for it. The complete dataset is utilized as input, but only the minority class is examined. KNN was utilized by SMOTE to discover new instances (or to produce fake data). In the vast majority of situations, it has no effect. The new instances don't just repeat minority cases that already exist. As an alternative, for each target class and its nearest neighbors, the computation applies component space tests to produce new models that merge objective case qualities with their highlights. The test scope widens as a result of this method, which makes highlights for each class more accessible [41].

E. FEATURE SELECTION METHOD

The primary focus of research in statistical science, data mining, and machine learning is feature selection. Several feature identification techniques have been deployed in recent years to healthcare datasets to obtain more useful information. On clinical datasets, feature selection techniques are used to predict DM, CVD, strokes, hypertension, thalassemia, and other chronic disorders. When the data contains more important and non-redundant qualities, different learning algorithms perform more effectively and produce more accurate results. Since medical datasets contain a huge number of duplicated and irrelevant attributes, finding fascinating disease-related components requires an effective feature selection technique [42]. Feature selection aims to select some important features and exclude less significant ones. Finding out the best subset of features that can improve the results is a big challenge. A thorough categorization model may be constructed by reducing the size of features and eliminating unneeded characteristics [43]. It is often preferable to remove inessential, as well as incomplete and inaccurate information before applying any model to the data in order to obtain more accurate results faster. Practically it is very critical to shrink the dataset's dimensionality. The primary problem of feature reduction is identifying the appropriate subset of traits to yield the best classification performance. Moreover, the complexity drops

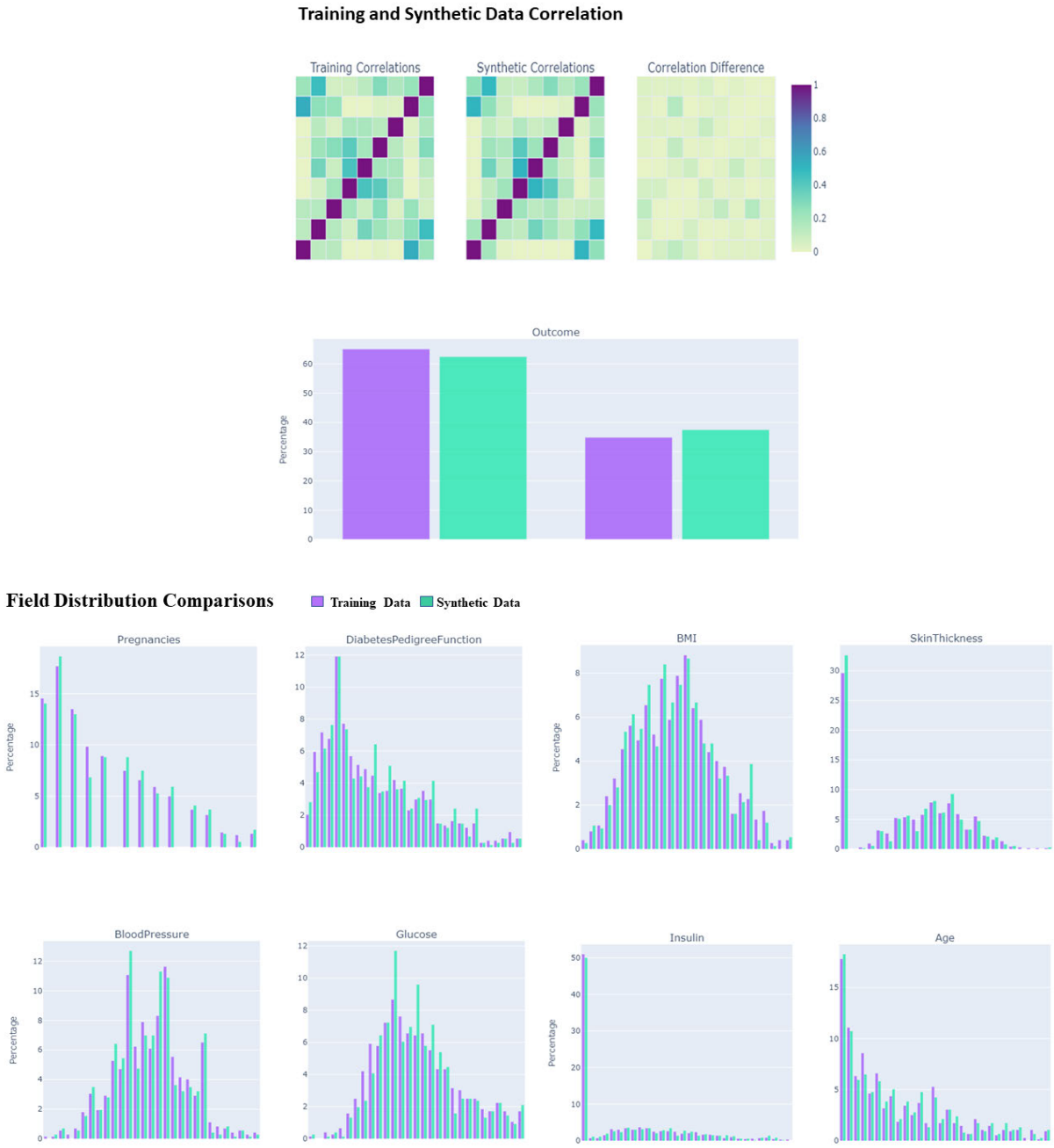


FIGURE 4. Report generated for synthetic data.

exponentially when the most crucial features are chosen [42]. This method shortens the training period of the learning algorithm, improves prediction accuracy, and makes data more comprehensible. It also makes data easier to visualize. There are three broad categories of typical machine learning feature selection strategies:

Filter Method, Wrapper Method and Embedded Method. Many authors are now opting for hybrid methods that combine both approaches, with somewhat same promising results [42].

1) FILTER METHOD

Unlike the following learning algorithms, filter approaches have a preprocessing stage. They use different approaches to choose features. A score or assessment criterion is used to choose a collection of qualities depending on how important each feature is to the target variable [44], [45].

2) WRAPPER METHOD

Wrapper Method is a feature selection technique that ranks a subset of traits based on how well a predictive

model that was trained alongside them predicted the future. A classifier that calculates the importance of a given subset of attributes is used for the evaluation. Although there is evidence that these strategies are effective, they are computationally expensive [46], which makes them less common [45], [47].

Boruta Method: Boruta is built on the same idea as the random forest classifier: random variations and relationships may be reduced by infusing randomness into the system and getting data from a collection of randomized samples. In this case, the additional randomization will allow us to determine which features are most important [48]. The steps of the Boruta algorithm are defined in Algorithm 1.

Algorithm 1 Boruta

1: Increase the size of the data by including duplicates of all variables (the information system is always extended by at least five shadow attributes, even if the number of attributes in the original set is lower than 5).
2: Rearrange the other characteristics to eliminate any connections with the outcome.
3: Run a random forest classifier using the enlarged information system, then gather the generated Z scores.
4: Determine the shadow attribute with the greatest Z score (MZSA), then hit every attribute that scored significantly higher than MZSA.
5: Use the MZSA to run a two-sided test of equality for each characteristic of unknown relevance.
6: If the importance of the qualities is significantly lower than MZSA, mark them as ''unimportant'' and permanently delete them from the information system.
7: Consider qualities to be ''important'' if their relevance is much higher than MZSA.
8: Eliminate all shadow properties.
9: Continue the process technique characteristics have been assigned a priority or the algorithm has used all available random forest runs, whichever comes first.

3) EMBEDDED METHOD

These techniques choose features during the learning process and are often given to the learner. By applying their various evaluation criteria during various stages of the search process, the current model also benefits from the two preceding models. Filter capabilities and Wrapper techniques are combined in embedded approaches. Algorithms with internal feature selection techniques carry out this [49]. Unlike the wrapper method, embedded techniques interact with learning algorithms at a lower computational cost. It retains feature dependencies and considers not just the relationship between input and output features, but also looks locally for traits that allow for stronger local discrimination. It selects the optimal subset for a given cardinality using independent criteria. The learning algorithm is used to select the best subset from those with changing cardinality [50].

F. HYPERPARAMETER TUNING

Each algorithm has some hyperparameter that can be used to control the process of learning and can be modified before training time. This setting of the hyperparameter is known as hyperparameter tuning. To get better performance of the ML model, it should be preferable to obtain optimal hyperparameters. Before model training, hyperparameter tuning is performed to get better results. All of the hyperparameters that will be modified in this study [51]are the number of epochs and batch size. Where Each approach produces the optimum deep neural network hyperparameter. To adjust hyperparameters, three basic strategies are used: grid search, random search, and Bayesian optimization.

1) GRID SEARCH

Grid Search is a conventional approach that evaluates all hyperparameter combinations. Grid Search employs the learning rate and several layers as hyperparameters. Initially, a set of values is specified for each hyperparameter. Each iteration estimates the hyperparameter combination. Finally, the optimum hyperparameter combination for the learning algorithm is chosen and executed [52]. The Grid Search considers all possible combinations in the hyperparameter space defined. In python the GridSearchCV function from the sklearn library is used to tune hyperparameters using Grid Search [51], [53].

2) RANDOM SEARCH

In this method, a combination of hyperparameters is selected randomly. So this is not considered the whole set of combinations of hyperparameters. In python, the RandomizedSearchCV function defined in the sklearn library is used to tune hyperparameters using random search [51], [53].

3) BAYESIAN OPTIMIZATION

Bayesian optimization is a method based on the Bayes theorem that selects the set of the optimal hyperparameters for the next evaluation by taking the previous evaluation into account. The number of hyperparameter combinations that Bayesian optimization will try is explicitly specified. In python the BayesSearchCV function defined in skopt library is used to tune hyperparameters using Bayesian optimization [51], [53].

G. GREY WOLF OPTIMIZATION ALGORITHM

In 2014, author [54] proposed Grey Wolf Optimizer (GWO) is a grey wolf-inspired meta-heuristic optimization algorithm. This algorithm mimics the leadership organization and hunting style of grey wolves in the wild. Wolves in this stage usually live in groups of 5 to 12 associates, with two serving as leaders. Grey wolves are categorized as alpha, beta, delta, or omega wolves focused on their effectiveness, decision-making skill, and efficiency [55], [56]. The leaders of the hunters, or ''alpha wolves'', decide on hunting tactics. They are the most dominating wolves in the pack since the others

must obey their orders and follow their lead. The pack's alphas don't have to be the strongest individuals, but they must be the most adept at leading the group as a whole. Wolves in the beta are second in the hierarchy. When making decisions, a beta wolf assists the alpha. The beta wolf replaces the alpha wolf if he passes away or ages. As one of the lowest levels in the hierarchy, they must uphold discipline and reaffirm the alpha's orders to the pack. Omega wolves are the lowest members of the hierarchy and are used as a convenient excuse. They should be the last to feed and should submit to the dominant wolves in the den. A subordinate wolf in the pack is referred to as a delta wolf, which is a wolf that is neither an alpha, beta, nor omega. Although they are subordinate to alphas or betas, delta wolves are in charge of omega wolves [57]. The steps of the GWO algorithm are defined in Algorithm 2

Algorithm 2 GWO

Initially Set the grey wolf population $X_i(i = 1, 2, \ldots, n)$
Initialize $a, A, andC$
Determine the fitness of each search agent.
Consider the best search agent as X_α
Consider the second best search agent as X_β
Consider the third best search agent as X_δ
while $(t < Maxnumberofiterations)$
for each search agent
Modify the current search agent's position.
end for
Modify $a, A, andC$
Evaluate the fitness of all search agents
Modify $X_\alpha, X_\beta andX_\delta$
$++t$
end while
return X_α

H. PROPOSED HYBRID METHODOLOGY

This section describes the hybrid hyperparameter optimization technique that we developed by combining Grid Search with Grey Wolf Optimization (GS-GWO). The main structure of proposed approach is depicted in Fig.5.

Grid Search application at the outset of the search process is a crucial objective of this hybrid technique. Grid Search subsets are assessed using a stopping criterion. The fitness function must either stay constant over many rounds or vary in a negligibly small way, and the current wolf performance metric must be compared to the most recent best learning measure for Grid Search to send the data to GWO. When the halting requirements are satisfied, the GWO algorithm takes over the search process. The GWO algorithm's starting population is made up of the best subsets discovered by the Wolf population. In our work, we have used GS-GWO method for identifying the hyperparameters of machine learning algorithms as illustrated in Table 6.

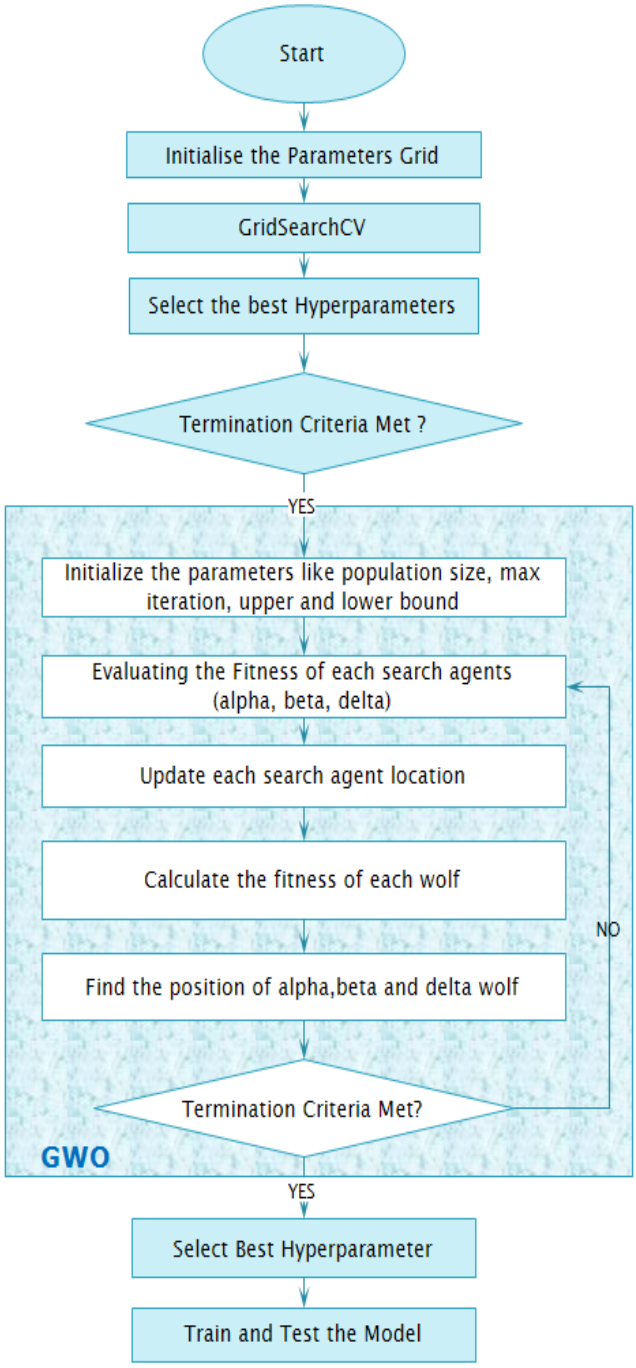


FIGURE 5. Architechture of proposed hybrid methodology.

I. CLASSIFICATION ALGORITHMS

1) K-NEAREST NEIGHBOUR

A machine learning algorithm whose working correlates with a lazy learner is called KNN, in which learning occurs since acquiring test data and even the method's training data have no time overhead. The test data formula will use the distance measure to determine which k data points are relatively close to each test data point, and a conclusion will be drawn based on the category data of those k data points. In the binary

TABLE 6. Hyperparameters of different ML Algorithms.

Algorithms	Hyperparameters
Logistic Regression	penalty='l2',class_weight='balanced', max_iter=100000, C=10, solver='liblinear'
Naive Bayes	type='multinomial', alpha=150
Decision Tree	criterion='gini', splitter='best', max_depth=16, max_features=15, min_samples_leaf=5, min_samples_split=0.0001
Random Forest	class_weight='balanced',criterion='gini', max_depth=9, max_features=17, min_samples_leaf=6, min_samples_split=30, n_estimators=32
K-Nnearest Neighbors	weights='distance', metric='minkowski', n_neighbors=16, leaf_size=10
SVM	C=10, gamma=0.1 , Kernel='Linear'
ANN	Batch size = 30,Epochs = 20,Dropout rate = 0.1,Learning rate = 0.001,Activation function= tanh, Kernel Initializer = uniform,No. of neurons in layer 1 = 16,No. of neurons in layer 2 = 4
MLP	lr= 0.1, max_epoch= 100, module_hiddenTwo= 10, optimizer_weight_decay= 0.1
CNN	filters=32, kernel_size=2, activation='relu'

classification problem, the test data is frequently the category with the highest percentage of the k sample points [39].

2) NAIVE BAYES

This technique is based on the Bayes theorem and in which each pair of attributes is independent of each other [58]. It works successfully and may be used in several real-world contexts, such as spam filtering, document or text categorization, and so on, for both binary and multi-class categories. The NB classifier may be employed to accurately detect noisy events in data and develop a credible prediction model. The key advantage is that it takes less training data than more advanced algorithms to estimate the relevant parameters swiftly [48]. However, owing to its heavy assumptions on feature independence, its performance could be hampered. The most common NB classifier improvements are Gaussian, Multinomial, Complement, Bernoulli, and Categorical.

3) RANDOM FOREST

An ordinary bagging algorithm is Random Forest (RF) [39]. Each classifier is trained by RF using a randomly selected section of the dataset and a randomly selected subset of the features in contrast to traditional decision trees. The outcomes of each trained classifier's prediction are different for the same input. The final prediction result is determined by voting on each trained classifier's output, often using the plurality or mean. As the features are scattered at random, the method will increase the variety of its own classifiers, which will improve the model's prediction performance.

4) DECISION TREE

It is a tree-structured based supervised machine-learning algorithm. This tree is made of nodes and leaves. The whole set of choices or effects is reflected in the leaves. The data are divided into smaller parts at decided nodes. To compile the necessary data for making decisions, a decision tree is constructed using a variety of techniques, including ID3, CART, and C4.5 [59].

5) SUPPORT VECTOR MACHINE

Another famous supervised machine learning approach is Support Vector Machine (SVM) which can work as a classifier and a regressor. This algorithm creates a decision boundary that can separate an N-dimensional space into classes so the new data points can be easily put in to correct class. These data points are called Support Vectors while the best decision boundary is called hyper-plane. The margin between the hyper-plane and data points should be as maximum as possible so the data points are categorized correctly [41].

6) LOGISTIC REGRESSION

Another form of supervised learning technique is Logistic Regression (LR). The model is statistical. Logistic regression forecasts the likelihood of the target value. The target characteristic is separated into two categories: success and non-success. It produces 1 in case of success and 0 in case of failure.

$$P = \frac{1}{1 + e^{b_0 + b_1 \times x + b_2 \times x^2}} \tag{1}$$

In eqs (1) P is the predicted value, b0, b1, and b2 are biases, and x is a variable that reflects a logistic regression. It is utilized in a variety of social science and medical machine learning applications, such as spam identification, diabetes diagnosis, cancer detection, etc. [41].

7) ARTIFICIAL NEURAL NETWORK

ANN is a sort of machine learning that mimics how the human brain works. ANN is meant to learn from input and can categorise and anticipate output, similar to how neurons in the human brain process and respond to information. An ANN has a data input layer, hidden layer(s), and output layer, as well as many nodes that operate as neurons. ANN are non-linear statistical models that can solve complicated issues. However, the effectiveness of an ANN for prediction is heavily dependent on choosing the right parameters and activation function [60], [61], [62], [63], [64].

8) MULTI LAYER PERCEPTRON

The layers of a neural network include hidden input and output layers. The input layer accepts the data, and the output layer provides the results. A hidden layer exists between the input and output layers. The body's neural network inspired the neural network. Network neurons, like human neurons, display probabilistic behavior. In neural networks, processing time is substantially longer. Multilayer Perceptron artificial neural networks increase complexity and density by allowing for a large number of hidden layers between the input and output layers. Every node on a certain layer is linked to every node on the following tier. Multilayer Perceptron models [65]

are therefore fully linked networks that may be used for deep learning.

9) CONVOLUTIONAL NEURAL NETWORK

A CNN is a type of multi-layer perceptron similar to a standard neural network in which each neuron receives distinct inputs. These self-learned neurons learn from data using weight and bias by performing operations like the dot product. CNN is composed of several layers, including a convolutional layer, a maximum pooling layer, a flattening layer, and a fully connected layer. The convolutional layer's purpose is to learn the feature representation for the incoming data. It is the network's heart, with local connections and weights for common properties. In the first step, input parameters are routed through the kernel, and outputs are routed through a nonlinear activation function ReLU, which does not activate all neurons at the same time. It exclusively activates neurons with values between 0 and 1. The pooling layer, which may be regarded of as a fuzzy filter since it reduces the dimensionality of the features while enhancing their robustness, is then applied to the output neurons. Finally, the completely linked layer receives signals from the preceding layers and sends them to all of the neurons in the system. The classification is subsequently performed by the output layer, often a softmax classifier [66].

J. PERFORMANCE EVALUATION

For evaluating the performance and efficiency of the proposed model, a binary matrix is used which is called the Confusion matrix as shown in Table 7. And with the help of the Confusion matrix, different evaluation metrics such as Accuracy, Precision, Recall, and Matthews correlation coefficient (MCC) are used. The following measures were considered in this research to analyze the performance of the methodologies used [43], [67], [68].

TABLE 7. Confusion matrix.

		Predicted Value	
		Negative (0)	Positive (1)
Actual Value	Negative (0)	TN	FP
	Positive (1)	FN	TP

Where

•True Positive (TP) - Positive instances that are correctly classified as positive outputs.

•True Negative (TN) - Negative instances that are correctly classified as negative outputs

•False Positive (FP) - Negative instances that are incorrectly classified as positive outputs

•False Negative (FN) - Positive instances that are incorrectly classified as negative outputs

All of the evaluation metrics are written down in the form of eqs(2)(3)(4)(5)(6)(7):

1) ACCURACY

Classification accuracy is defined as the ratio of correct predicted values to total predicted values and is mathematically represented as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

2) PRECISION

Precision is sometimes referred to as favorable predictive value (PPV), is expressed as the ratio of correct predictions to total correct values, which includes both true and false predictions, and is represented mathematically as follows:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

3) RECALL

The recall, sensitivity, or true positive rate (TPR) is expressed mathematically as the ratio of accurately predicted values to the sum of correct positive predictions and incorrect negative predicted values:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

4) MATTHEWS CORRELATION COEFFICIENT (MCC)

The value of MCC ranges between 1 and -1. The -1 value of MCC denotes total conflict between prediction and observation, whereas 1 represents exact prediction and 0 represents random prediction.

$$MCC = \frac{(TP \times TN)(FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

5) ROC-AUC-SCORE

Prediction scores are used to compute the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). The area under the ROC curve (AUC) represents a test's ability to distinguish diagnostic groups or classes. The AUC value ranges from 0 (the classifier incorrectly diagnosed all classes) to 1 (perfect diagnostic performance between classes). The Receiver Operating Characteristic (ROC) curve is a graphical representation of classifier performance. The ROC curve compares the true positive rate (sensitivity) to the false positive rate rate (1-specificity). Furthermore, AUC is the area under the ROC curve. The higher the classification accuracy, the closer the AUC value is to 1.

$$ROC = \frac{Sensitivity + Specificity}{2} \tag{7}$$

K. STACKED GENERALIZATION

Stacking is an ensemble learning technique to combine multiple classification models by a meta-classifier. This is also called Stacked Generalization or Stacking Classifier. The method's fundamental idea is to create a more powerful meta-model (level 1 models), that combines predictions from several base learners (level 0 models), to reduce generalization error [58], [68], [69], [70].

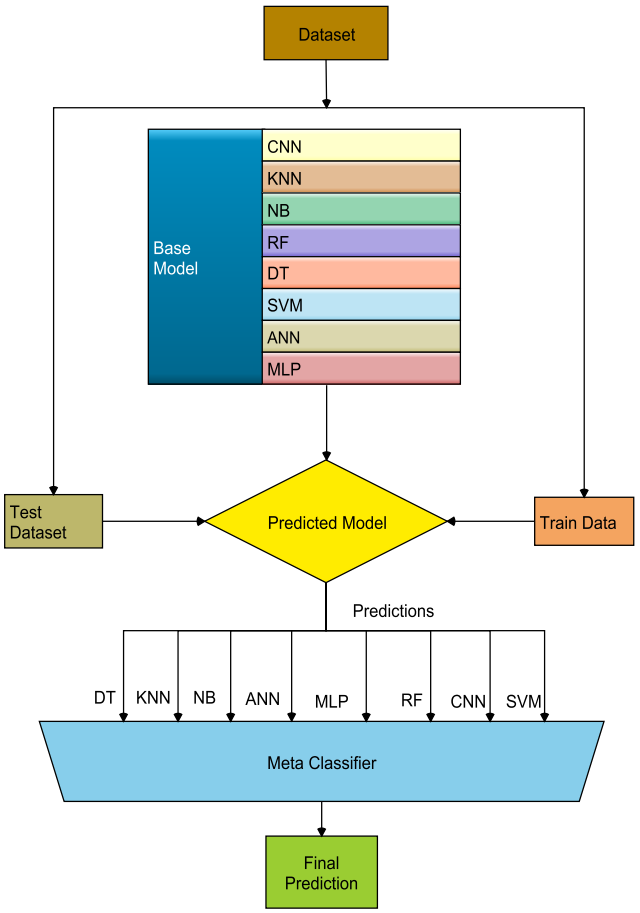


FIGURE 6. Architecture of stacking classifier.

TABLE 8. Environmental Setup.

CPU	Intel (R) Core ™ i5 10300H@ 2.50GHz
RAM	8GB
GPU	4GB
Software	Anaconda/ Python3

TABLE 9. Step by step Performance Evaluation of ML Algorithms on PIMA dataset.

Performance Measure/ Algo	Accuracy	ROC AUC Score	Matthews Corr Coef	Precision	Recall	F1-Score	Error Rate
(A) After Feature Scaling							
KNN	77.07	71.14	40.43	79	91	85	22.93
NB	77.72	72.07	41.1	80	91	85	22.28
RF	79.67	73.35	46.69	82	93	87	20.33
DT	70.58	64.8	25.99	76	84	80	29.42
SVM	80.32	73.64	47.9	81	96	88	19.68
LR	80.97	75.85	50.02	83	93	88	19.03
(B) After SMOTE							
KNN	91.875	73.48	37.96	83	82	82	8.125
NB	78.75	79.4	54.1	87	89	88	21.25
RF	100	79.33	53.6	87	87	87	0
DT	100	75.12	45.81	84	86	85	0
SVM	90	73.63	42.2	83	83	83	10
LR	83.5	76.9	50.13	86	84	85	16.5
(C) After Boruta							
KNN	91.275	78.85	49.5	87	94	90	8.725
NB	88.75	83.27	61.08	89	98	93	11.25
RF	100	83.55	58.81	90	96	93	0
DT	100	77.35	45.75	86	94	90	0
SVM	87.85	80.07	56.18	95	96	95	12.15
LR	87.36	79.57	54.57	87	95	91	12.64
(D) After Grid Search							
KNN	94.56	91.57	64.5	93.24	92	92.61	5.44
NB	88.75	89.27	72.08	96	98	97	11.25
RF	96	94.55	64.81	92	96	93.9	4
DT	96.23	97.8	69.75	92	97	94.4	3.77
SVM	94.67	98.4	70.18	93	96	94.4	5.33
LR	87.36	86.57	54.57	85	96	90.1	12.64
(E) After GS-GWO							
KNN	97.6	91.57	64.5	94.3	95	94.65	2.4
NB	89.5	89.27	73.7	98	99	98.5	10.5
RF	97	94.55	68.1	95	97	96	3
DT	97.32	97.8	69.5	94	96	95	2.68
SVM	96.7	98.4	75.18	97	98	97.5	3.3
LR	88.4	87	53	86	97	91.5	11.6

Fig.6 demonstrates the stacked generalization schemes. Stacking is a two-step procedure in Machine Learning. In the first phase, we use a initial-level classifier, and at the second phase, we apply a meta-classifier that learns from the prediction made by the first-level classifier. The free choice of base learners is an important feature of stacked generalization. So this is different and even better than traditional ensemble methods. The individual classification model is trained using the whole training set, and the meta classifier is fitted using the initial-level classifier's outputs. There are two ways of training the meta-classifier: 1) The outputs of the initial level classifiers are used as inputs or features to the meta-classifiers. 2) The basic learners' predicted probabilities are utilized to develop the second-level model (final model). Furthermore, the outcomes of base classifiers may be complementary, and this arrangement may be beneficial in enhancing the final meta-model's performance [59], [71].

In this paper, for building a stacking classifier, eight classifiers NB, KNN, RF, DT, SVM,ANN, MLP,CNN are used at level 0, while at level 1, Logistic Regression is used. The level 0 classifiers' prediction output is sent to the level 1 classifier.

IV. EXPERIMENTAL ENVIRONMENT AND RESULTS

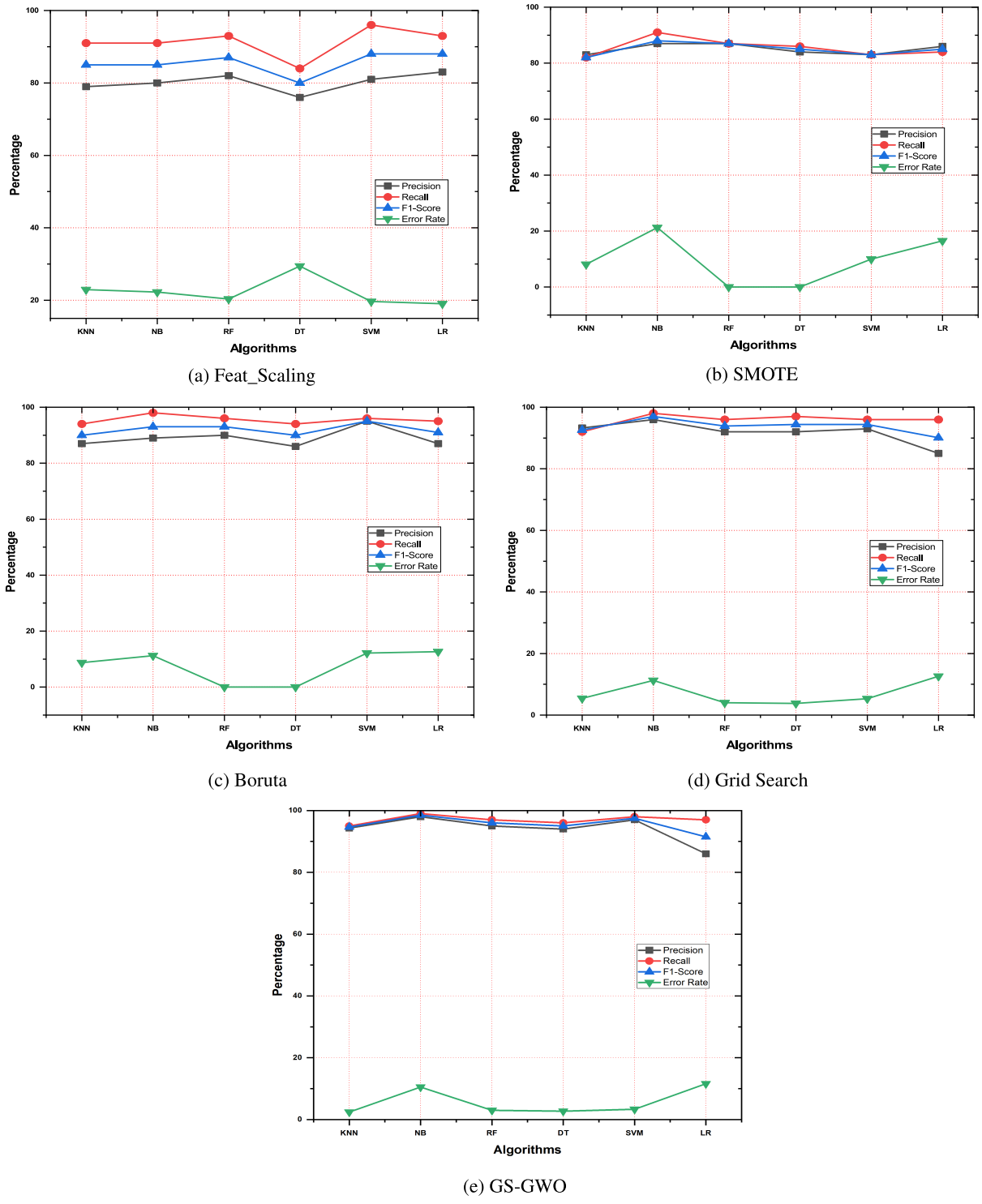
The proposed system was developed in a variety of situations. The environment setup of the developing system is shown in Table 8.

After preprocessing of data, we apply feature scaling on the data set,construct the model and evaluate its performance. After applying feature scaling and the performance is represented in the form of Table 9 (A). Graphical representation of Table 9(A) is shown in Fig. 7a. Next, we apply the Balancing algorithm SMOTE on the data set, again train the model, and observe the model's performance after applying SMOTE. The performance is expressed in the form of Table 9(B). Graphical representation of Table 9(B) is shown in Fig. 7b.

Next, we apply Feature Selection methods like filter methods: variance methods, correlation coefficient, information gain, chi-square methods, and wrapper method: Boruta on the data set to select important features.

In which Boruta gave better output. Train the model again and examine the model's performance after applying Boruta, and the performance is recorded in the form of Table 9(C). Fig. 7c exhibit a graphical depiction of Table 9(C).

Next, we apply the Grid Search i.e. Hyperparameter tuning algorithm, train the model then examine the model's performance after applying Grid Search which is represented



(a) Feat_Scaling

(b) SMOTE

(c) Boruta

(d) Grid Search

(e) GS-GWO

FIGURE 7. Step by step performance evaluation of ML Algorithms on PIMA dataset.

in the form of Table 9(D). Graphical representation of Table 9(D) is shown in Fig. 7d.

Next, we apply the proposed hybrid GS-GWO hyperparameter optimization technique GS-GWO, Train the model again and observe the model's performance after using GS-GWO, and the results are shown in Table 9(E) Graphical representation of Table 9(E) is shown in Fig.7e.

However we also apply some advanced model like ANN, MLP, CNN along with these traditional machine learning algorithms. As a result, MLP, ANN, and CNN have been

TABLE 10. Performance of some advance algorithm on PIMA dataset.

Performance Measure/ Algo	Accuracy	ROC AUC Score	Matthews Corr Coef	Precision	Recall	F1-Score	Error Rate
After GS-GWO							
MLP	80.3	80.5	79	81	85	83	19.7
ANN	85.5	83	81	84	86	85	14.5
CNN	89	87.8	86	91	95	93	11

TABLE 11. Performance of algorithms on PIMA dataset.

Algorithms	Accuracy	ROC_AUC Score	MCC	Precision	Recall	F1-Score
KNN	95.3	88.2	68.5	94	92	93
NB	86.5	87.1	62.1	96	98	97
RF	95	86	64.1	97	97	97
DT	96.4	85.2	60.5	93	94	93
SVM	94.7	88.4	70.18	95	97	96
LR	85.4	79	71	90	96	93
MLP	78.3	80.5	62	89	85	87
NN	76.2	84.4	72.4	88	86	87
CNN	87	82	63.3	89	95	91

TABLE 12. Performance of algorithms on ADRC dataset.

Algorithms	Accuracy	ROC_AUC Score	MCC	Precision	Recall	F1-Score
KNN	92.3	84.6	68	90	98	93
NB	95.4	89.8	78.5	95	98	96
RF	98.1	86.3	72.2	98	99	98
DT	98.6	87.2	75	98	100	99
SVM	97.7	85.4	69	97	99	98
LR	91.3	87.7	75.1	95	99	97
MLP	89.2	79.3	63.1	87	85	86
NN	87.5	75	62	89	91	90
CNN	91	80	64	88	88	88

TABLE 13. Performance of algorithms on FHD dataset.

Algorithms	Accuracy	ROC_AUC Score	MCC	Precision	Recall	F1-Score
KNN	92	87	68	95	98	96
NB	95.52	87.1	70.1	95	98	96
RF	97.5	89	74	97	97	97
DT	96.8	83.7	66.2	92	96	94
SVM	91.8	83	67	91	95	93
LR	97.7	87	70.1	97	99	98
LP	88	84	68	85	88	86
NN	91	85	68	87	86	86
CNN	92	84.6	70	90	95	92

utilized as classifiers in this diabetic illness prediction model. Table 10 summarizes the findings of studies on MLP, NN, and CNN.

It is worth noting that MLP and NN obtain less than 85% F1-score, whereas CNN achieves less than 94% F1-score. We also validate the performance of each model by using synthetic dataset based on PIMA dataset and the results are shown in Table 11.

We also used ADRC and FHD datasets, and the performance of each algorithm is depicted in the form of Table 12 and 13.

Next, we apply a stacking classifier, the proposed stacking classifier model is then trained and evaluated on different samples, namely the source and target datasets. In this study, the proposed diabetes prediction model is first worked on a larger open-source Pima Indian dataset then worked on ADRC dataset and FHD dataset. Next, we apply a stacking classifier, train the model and observe the performance of the model and get a 98.7% F1-score. Table 14 shows the performance of the proposed stacking classifier on different diabetes and Fig. 8 shows the graphical representation of variation in the parameters of performance among different datasets.

To assess model performance, we performed a statistical t-test between Stacking Classifier and KNN,NB,RF,DT,SVM,

TABLE 14. Performance of proposed Stacking Classifier on different dataset.

Dataset	Accuracy	ROC_AUC Score	MCC	Precision	Recall	F1-Score
PIMA	97	92.5	74	98.5	99.2	98.84
Synthetic data based on PIMA	96.33	92.8	74.3	98.20	97.12	98
ADRC	96.7	90.4	67.5	96.03	98.79	97.3
FHD	95.2	91.3	69.4	95.12	97.32	96.20

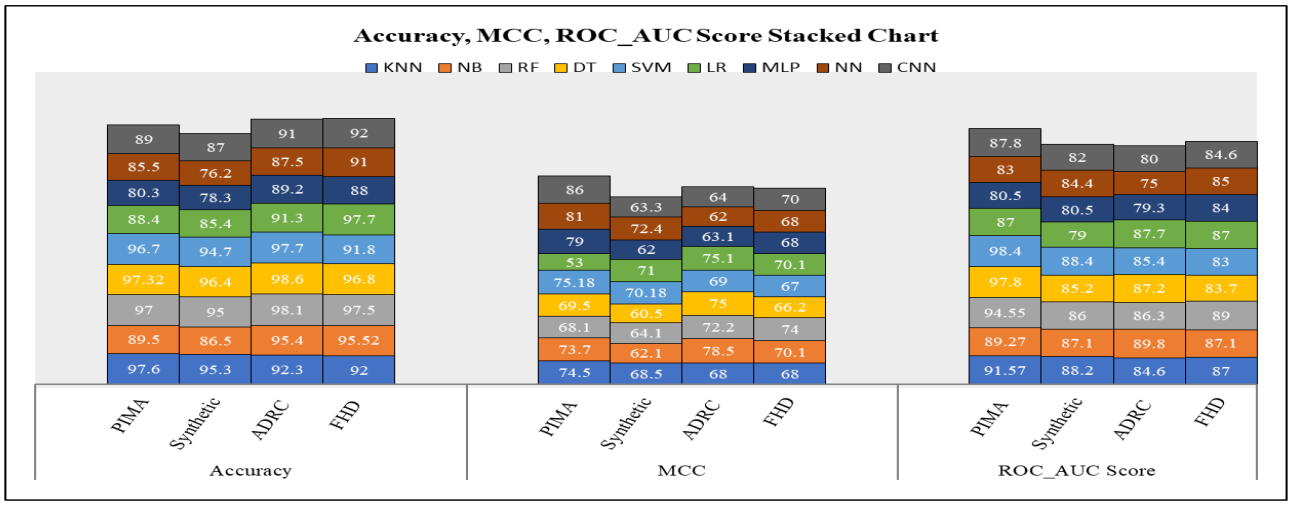
TABLE 15. P-Value after conducting t-test.

Dataset	Parameter	Proposed Model	KNN	NB	RF	DT	SVM	CNN
PIMA	Mean Accuracy	96.856	90.73333	94.93333	91.93333	94.4	91.66667	89.73333
	P- Value	-	0.0008	0.0004	3.65E-06	1.45E-07	0.004249	6.33E-11
ADRC	Mean Accuracy	94.42	89.8	92.6	93.13333	93	94.2	85
	P- Value	-	1.59E-08	0.023621	0.006823	0.008436	0.026186	1.19E-14
FHD	Mean Accuracy	96.856	92.06667	91.73333	91.26667	92.33333	91.13333	88.46667
	P- Value	-	4.47E-07	3.36E-08	5.35E-09	5.85E-11	2.31E-07	1.29E-10

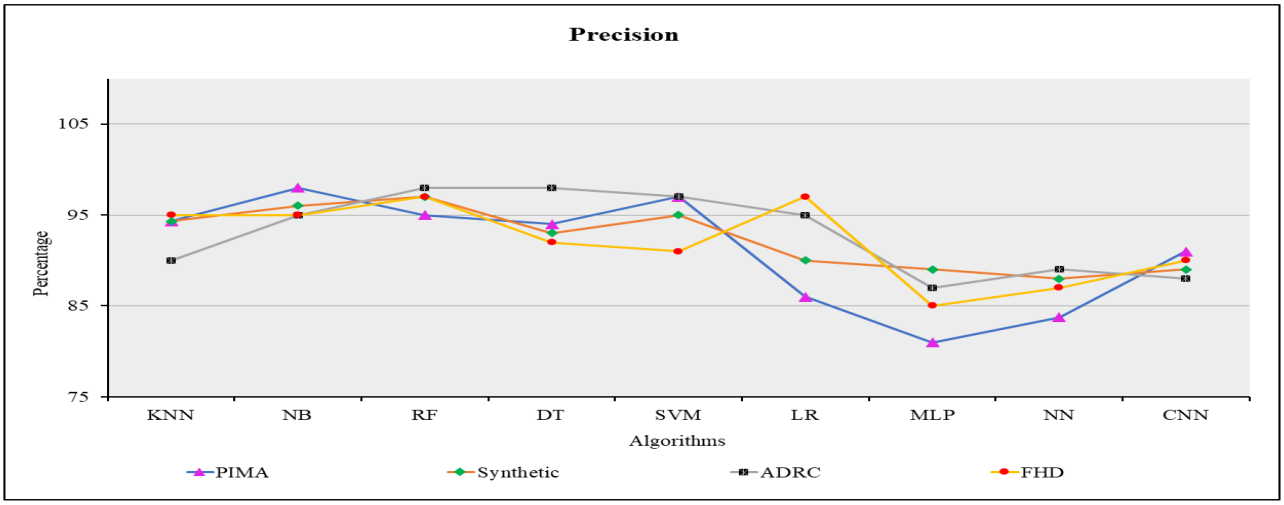
and CNN. We utilized the variance estimate by examining the dataset's dependence and computed the p-value. The null hypothesis suggested that there was no statistical difference in the model performances. However, an alternate hypothesis we explored is that there may be a difference in model performance. If the p-value is smaller than the significant value, the null hypothesis is rejected and the alternative hypothesis is accepted. There was a considerable variation in model performance. This test is carried out using the Scipi library, with a crucial statistical significance of =.05: The t-test is used to analyze six algorithms (executed for each instance independently 15 times) based on their F1 Score. The computation of the p-value is shown in Table 15, where it is clear that the p-value is smaller than the significance value i.e.,.05. Because the p-value is smaller than the significance value, so we rejected the null hypothesis and concluded that the proposed model's performance is distinct from others and superior in terms of F1-Accuracy when compared to other models.

V. MODEL INTERPRETATION

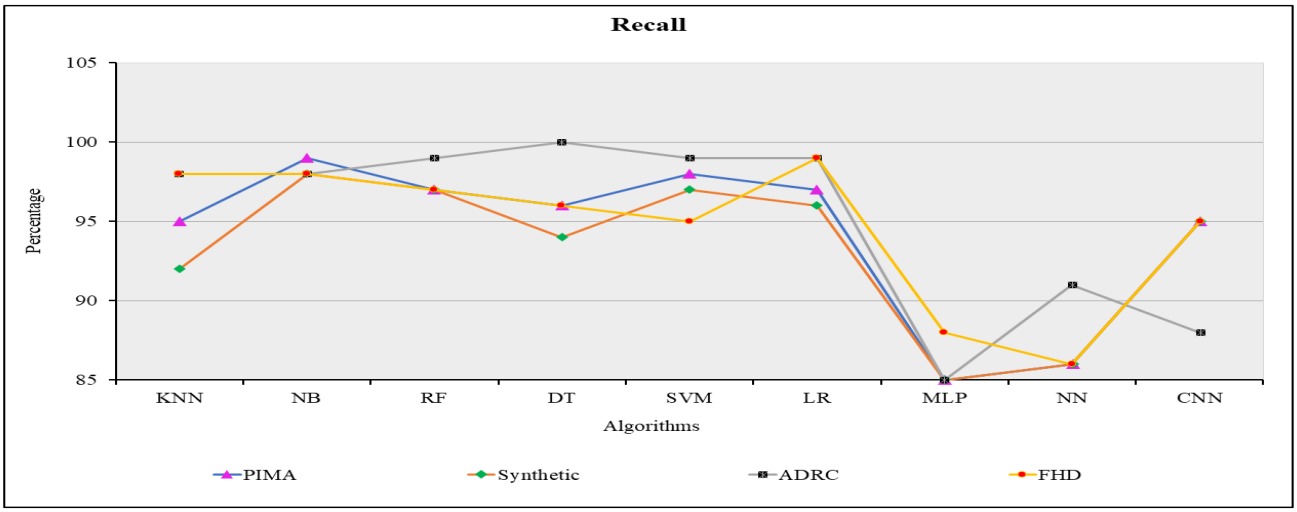
The depth to which the link between cause and effect can be identified inside a system is called interpretability. In another way, It is the degree to which a model can anticipate what will happen when input or computational parameters are changed. Interpretability or Explainability [72] is one of the most contentious issues surrounding the use of artificial intelligence (AI) in healthcare. As a result, from a medical standpoint, not only clinical validation but also explainability are vital in the clinical scenario. Explainability permits the resolution of conflicts between an AI system and human specialists, regardless of who is at fault. It should be highlighted that this will work best in circumstances of systematic mistakes, such as AI bias, rather than random error. Random mistakes are far more difficult to detect and will most likely go unexplored if the tool and the physician agree or will lead to disagreements between the tool and the physician. The findings of explainability tests are often expressed graphically or through natural language explanations. Both demonstrate to professionals how several factors influenced the ultimate recommendation. In other words, explainability can help doctors analyze system suggestions based on their expertise and clinical judgment [73], [74],



(a) Comparision based on Accuracy, MCC and ROC_AUC Score

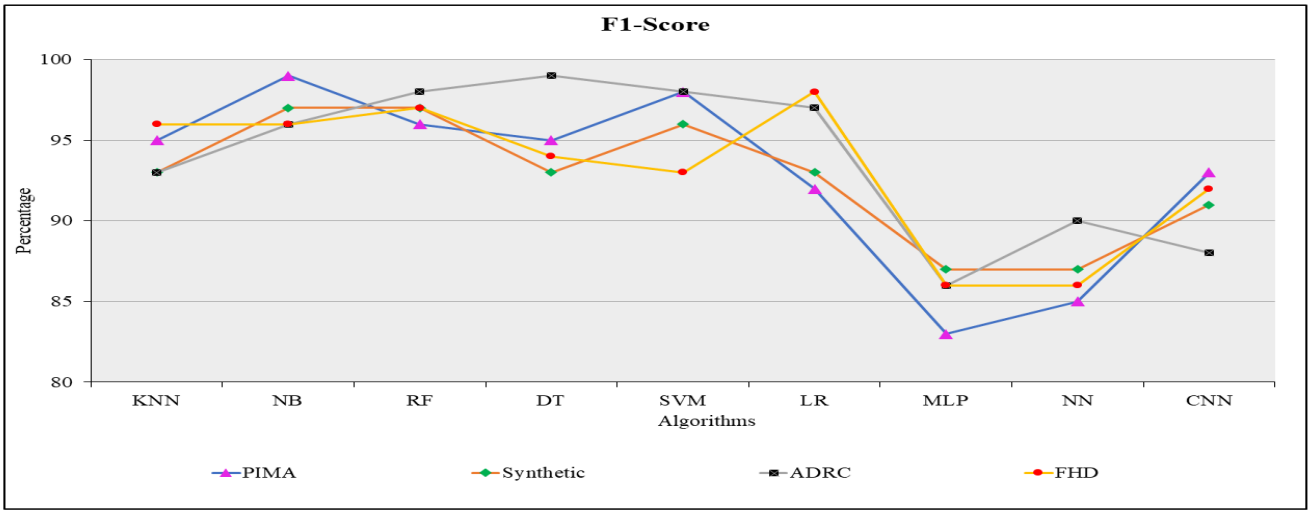


(b) Comparision based on Precision

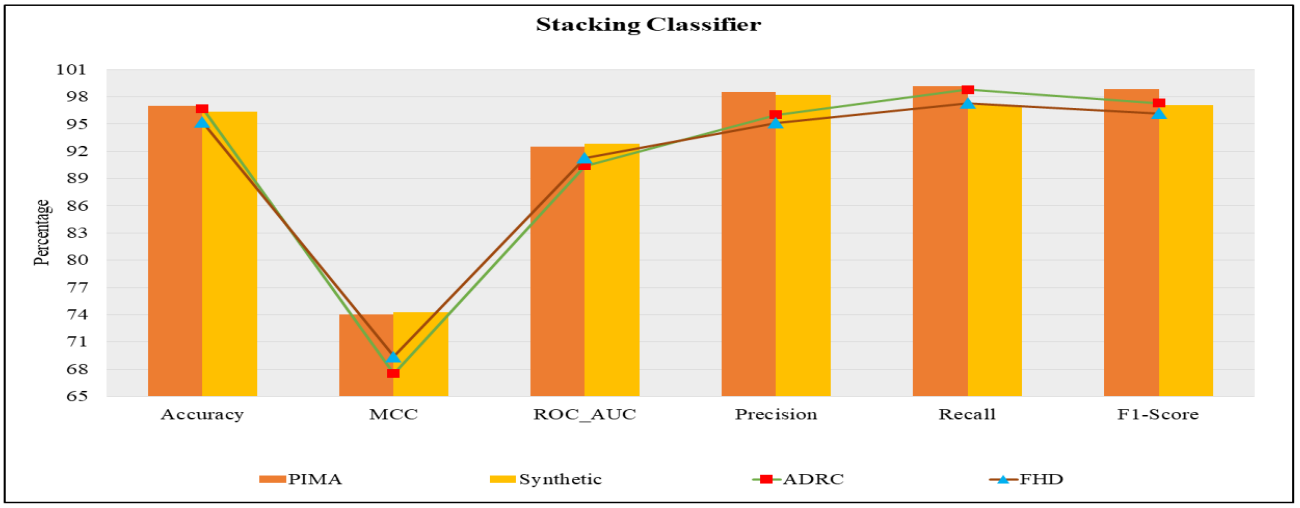


(c) Comparision based on Recall

FIGURE 8. Performance analysis of optimized proposed methodology with respect to different dataset.



(d) Comparision based on F1-Score



(e) Performance of Stacking Classifier

FIGURE 8. *(Continued.)* **Performance analysis of optimized proposed methodology with respect to different dataset.**

[75]. In this paper, two explainable algorithms are used LIME(Local Interpretable Model-Agnostic Explanations) and SHAP(SHapley Additive exPlanations) Let's look at the output variable in case of PIMA dataset. We can see in Fig. 9 that the seventh observation test data set has a value of 1, indicating that it is diabetes +ve. X_test[7] i.e. array([2.7187125 (Preg), 0.2406085 (Glu), -0.260103 (BP), 0.35257475 (ST), -0.41776815 (Ins), 1.08416645 (BMI), 0.29332509 (DPF), 0.91546889 (Age)])

These values come after the feature scaling operation perform on the dataset to scale the values of all attributes on the same scale. In Fig.9a the green color bar on the right side of the picture reflects support for positive diabetes, whereas the red color bar on the left side opposes the support. The variables BMI >.60 and glucose >-.15 strongly support positive diabetes for the chosen observation. In other words, at that instance X[7], combination of all values like BMI, glucose, no. of pregnancies and age was mostly responsible for +ve diabetes. In Fig.9b the local LIME model intercept is.3342, and the local LIME model prediction is 0.547 (Prediction_local). 0.85 prediction from the proposed model i.e. a stacking classifier. and it also visualizes the explanatory factors in order to determine how much they contribute. Similarly, for FHD dataset, the proposed model explainability, by using LIME algorithm, is represented in Fig.10 array([0.48419122 (Age), −1.06542721 (Gender), −0.64937829 (Glucose), −0.40071252 (SBP), −0.25695143 (DBP), 0.74234046 (Weight)]).

In Fig.10a, the green color bar on the right side of the picture reflects support for positive diabetes, whereas the red color bar on the left side opposes the support. The variables SBP<=−.35, DBP<=−.26 and glucose <=−0.57 strongly support negative diabetes for the chosen observation. In other words, a combination of all values like SBP, DBP and

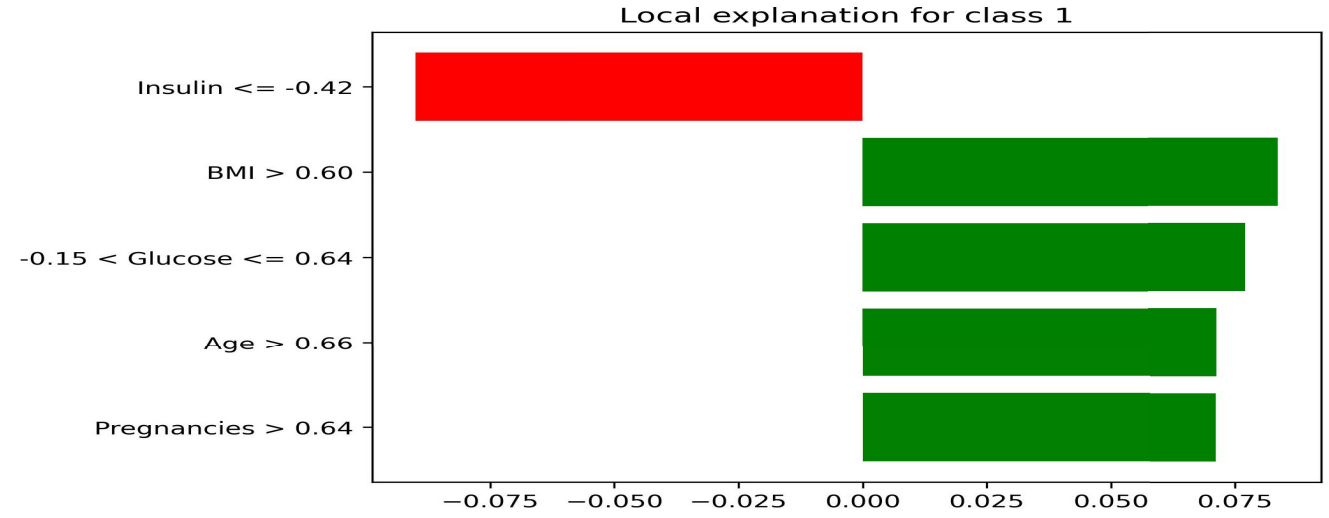
TABLE 16. Comparison of our approach with other existing research for prediction of chronic disease.

S. No.	Ref	Year	Dataset	Feature Scaling	Balancing Algo	Feature Selection Methods	Cross Validation	Hyper-Parameter Tuning	ML Algo	Work	Result (%)
1	[76]	2020	PIMA	✓	x	Corr	K-fold	✓	KNN, DT, RF, AdaBoost (AB), NB and XGBoost (XB), Multilayer Perceptron (MLP)	**Ensemble Classifier (AB+XB)**	P=84.3 R=78.7 F1=81.4
2	[40]	2021	UCI	x	SMOTE	FCBF	10-fold	x	**RF**,LR, KNN, NB, SVM	—	A=97.81 P=99.32 Specif =98.86
3	[77]	2021	Primary	x	x	Anova, Chi-Square, Recursive Feature Selection	K-fold	✓	LR, SVM, XB ,RF,	**RF**,Ensemble Classifier	A=73 P=74 R=73 F1=74
4	[78]	2021	Primary	x	x	Wrapper method, Filter Method	x	x	Linear Regression, RF, **SVM**, Gaussian Process	—	20.58 (RMSE)
5	[65]	2022	Kashmir real	x	✓	✓	✓	x	LR, SVM, MLP,GB, DT ,**RF**	-	A=98 P=72.7 R=87.5 F=79.4
6	[67]	2022	PIMA	x	SMOTE, TOMEK	Corr	K-fold	x	LR, SVM, KNN, RF, NB, Gradient Boosting (GB)	**Voting Classifier (RF+NB+GB)**	A=81.7
7	[79]	2022	PIMA	x	x	PCA	x	x	**NB**, RF,J48DT	—	A=79.13 P=81.6 R=88.08 F1=84.71
8	[80]	2022	PIMA	x	SMOTE	x	x	✓	CNN, CNN-LSTM, Conv LSTM, **DeepLSTM**, DCNN,	—	A=99.6 P=95.4 R=94.6 F1=94.9
9	[81]	2022	PIMA	x	x	Corr, PCA, IG(Information Gain)	x	✓	MLP, KNN, DT,**RF**	—	A=79.8 R=79.8 Specif = 71.4
10(i)	Our Work	2023	**PIMA**	✓	SMOTE	Boruta	K-fold	**Hybrid GS-GWO**	DT, SVM, KNN, RF, NB, LR, MLP, ANN, CNN	**Stacking Classifier (LR, RF+DT +SVM+NB+KNN +MLP+ANN+CNN)**	98(F1-score)
10(ii)	Our Work	2023	**Validation by using Synthetic dataset based on PIMA**	✓	SMOTE	Boruta	K-fold	**Hybrid GS-GWO**	DT, SVM, KNN, RF, NB, LR, MLP, ANN, CNN	**Stacking Classifier (LR, RF+DT +SVM+NB+KNN +MLP+ANN+CNN)**	98.7 (F1-score)
10(iii)	Our Work	2023	**ADRC**	✓	SMOTE	Boruta	K-fold	**Hybrid GS-GWO**	DT, SVM, KNN, RF, NB, LR, MLP, ANN, CNN	**Stacking Classifier (LR, RF+DT +SVM+NB+KNN +MLP+ANN+CNN)**	97.3(F1-score)
10(iv)	Our Work	2023	**FHD**	✓	SMOTE	Boruta	K-fold	**Hybrid GS-GWO**	DT, SVM, KNN, RF, NB, LR, MLP, ANN, CNN	**stackingClassifier (LR, RF+DT +SVM+NB+KNN +MLP+ANN+CNN)**	96.2(F1-score)

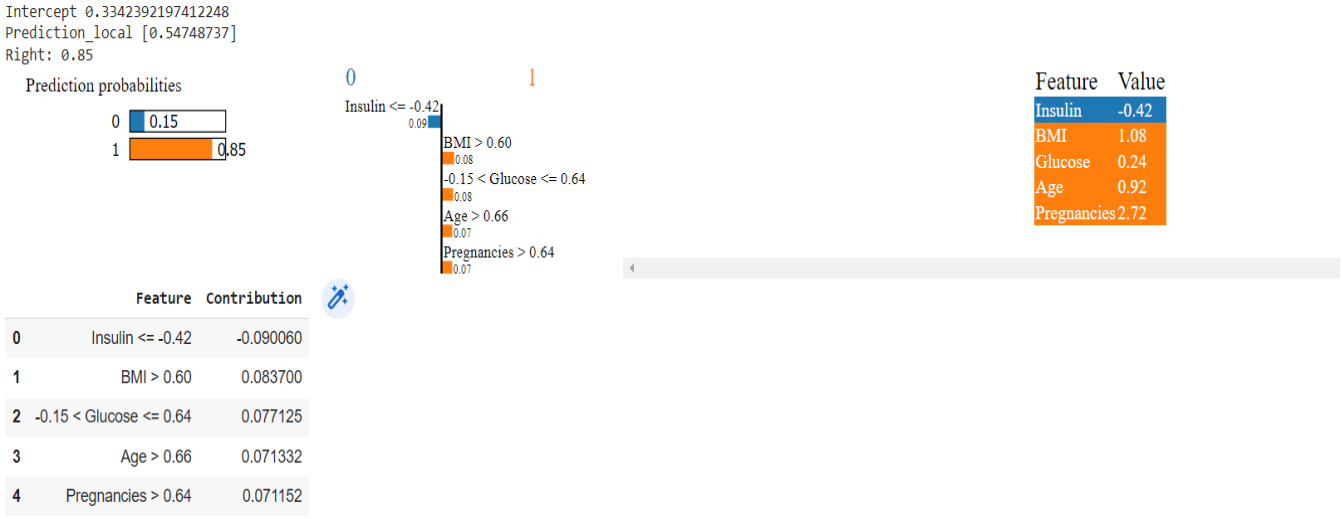
*Note: Bold, underlined approach represents the best result while 'x' represents not perform and '✓' represents perform

glucose of that instance was mostly responsible for -ve diabetes.

In Fig.10b, the local LIME model intercept is.4731, and the local LIME model prediction is 0.072 (Prediction_local).



(a) local prediction Bar_graph

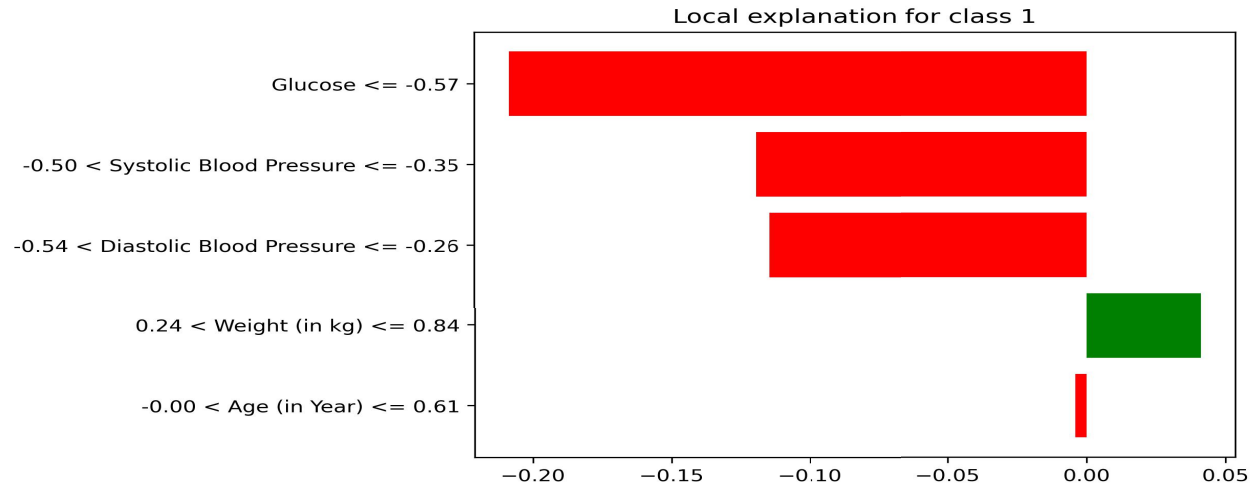


	Feature	Contribution
0	Insulin <= -0.42	-0.090060
1	BMI > 0.60	0.083700
2	-0.15 < Glucose <= 0.64	0.077125
3	Age > 0.66	0.071332
4	Pregnancies > 0.64	0.071152

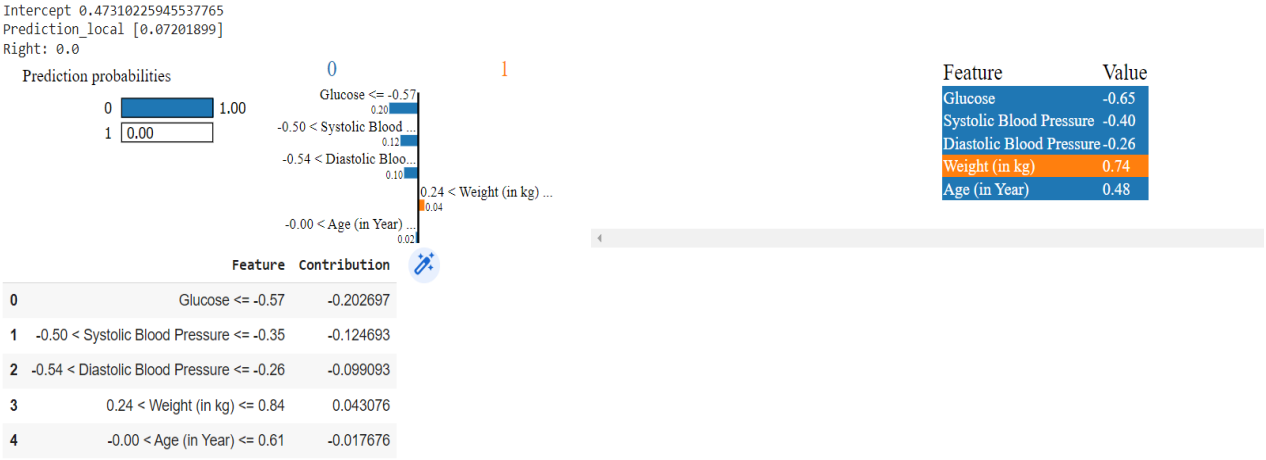
(b) Prediction explaination plot

FIGURE 9. **LIME prediction explaination for PIMA dataset.**

0 prediction from the proposed model, i.e., a stacking classifier means it has no diabeties in person. This also visualizes the explanatory factors to determine how much they contribute. For ADRC dataset, Fig.11a displays the feature significance determined from the average of absolute shapely values over the full dataset. For each dot: The vertical positioning indicates which feature is being shown. The color indicates whether that attribute was high or low for the particular row of data. The horizontal position indicates whether the influence of that value resulted in a greater or lower forecast. The six most important indicators for predicting diabetes disease are shown in Fig.11b, but their value varies. Alternatively, the global interpretation based on SHAP values shows that HbA1c, Age,Neuropathy and Insulin are the most relevant characteristics in proposed model. SHAP provides local interpretation for each sample and global interpretation of the entire dataset. Fig. 11c depict the categorization of a sample as high risk and low risk, respectively. The force plot depicts how each trait affects the risk categorization of each observation, as well as the direction and amount of the influence. The bar length represents the effect level for the associated characteristic. The preceding explanation depicts aspects that push the model output from the base value to the model output. Features that influence the forecast are highlighted in red, while those that influence the prediction are represented in blue. In the context of classification, red characteristics push the classification to be in the high-risk or +ve diabetes. At the same time, blue features indicate the prediction to be in the low-risk category or −ve diabetes. The base value is 0.615, and the anticipated value is.93. HbA1c = −.737 has the most influence on predicting positive diabetes, whereas



(a) local prediction Bar_graph



	Feature	Contribution
0	Glucose <= -0.57	-0.202697
1	-0.50 < Systolic Blood Pressure <= -0.35	-0.124693
2	-0.54 < Diastolic Blood Pressure <= -0.26	-0.099093
3	0.24 < Weight (in kg) <= 0.84	0.043076
4	-0.00 < Age (in Year) <= 0.61	-0.017676

(b) Prediction explaination plot

FIGURE 10. **LIME prediction explaination for FHD Dataset.**

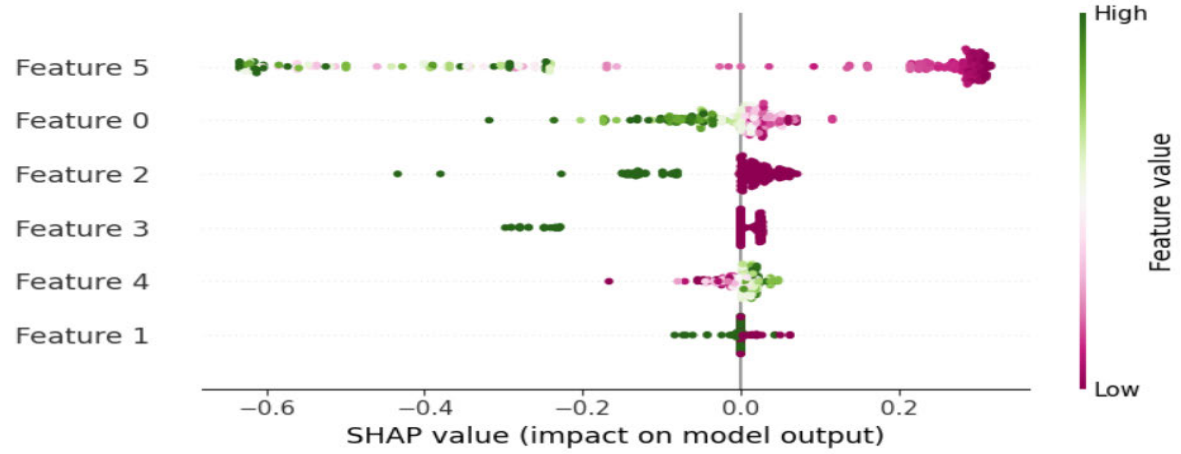
Hb feature has the greatest impact on predicting negative diabetes.

VI. COMPARISON

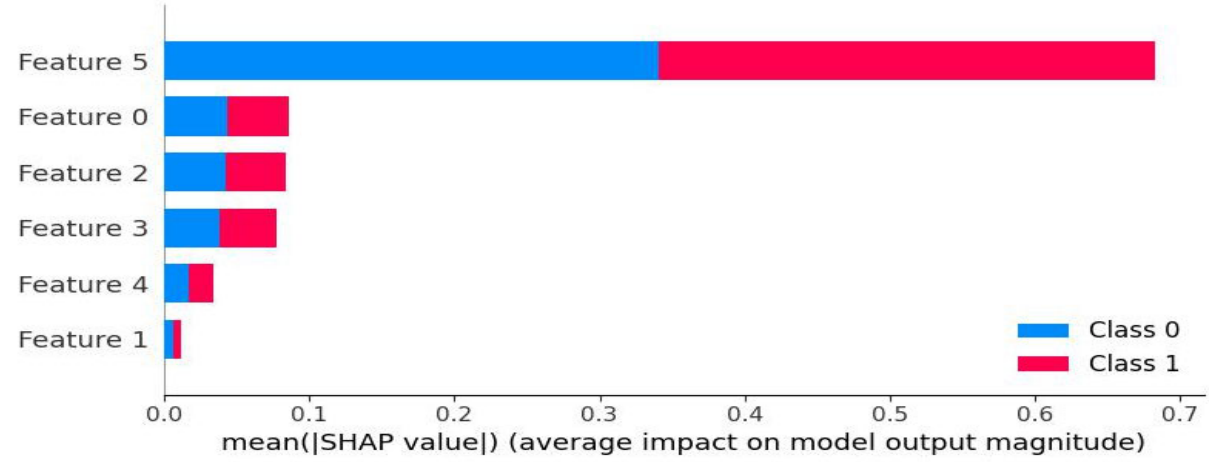
The performance of any prediction model can be influenced by a few aspects, such as the kind and size of the dataset, the algorithms and parameters of performance metrics, the distinct ideas used for that classifier, and the metrics of each classifier. Various strategies for the detection of diabetes have been offered in the past; results and comparisons of our work with existing studies and methodologies are shown in Table 16.

VII. DISCUSSION

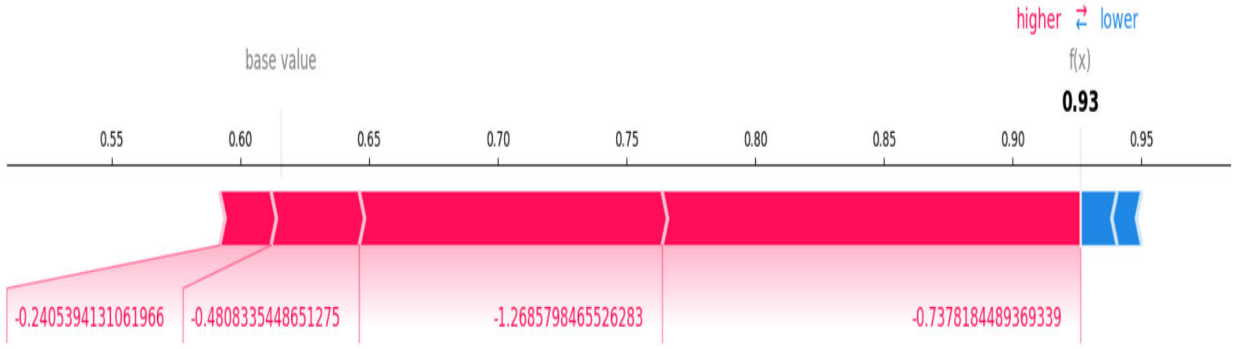
Unlike previous investigations, this work used three datasets and an ensemble ML approach to create prediction models. After the implementation of hyperparameter tuning, the performance of different classifiers for PIMA dataset was compared in which NB and RF classifier out of DT,KNN, SVM, LR, ANN, MLP and CNN Classifier, gave the best result with 97% of F1-Score. Additionally, after applying the stacking classifier model, the performance is increased by 1% and it was found to be 98% F1-Score. We can reduce the processing time by making better use of the GWO method. Then, use the proposed hybrid hyperparameter optimization approach GS-GWO on several ML classifiers, in which NB gave 98.5% F1- Score. Its findings are superior to those of basic Grid Search approach, and with 98.84% F1-Score after using a stacking classifier. But once applied synthetic data set for the validation of model performance, then NB and RF both gave 97% F1 score, and the stacking classifier gave 98% F1-Score, which is quite similar to the previous result. In addition, the real and unknown medical data utilized to train the models, as well as their influence on prediction accuracy and F1-score, were given in Table 14. On ADRC dataset the stacking classifier performed with 97.3% F1-Score, and on FHD dataset the model performed with 96.2% F1-Score. The experimental findings demonstrated



(a) Prediction explaination plot



(b) Prediction explaination plot



(c) local prediction Bar_graph

FIGURE 11. **SHAP prediction explaination for ADRC dataset.**

that the created prediction models, i.e., the suggested stacking classifier, outperformed alternative models. However, there is a performance difference between the proposed model and individual models, although it is minor difference in the test data. Which has been evaluated after applying a statistical t-test. The model and its interpretation can be useful for prac-

titioners in clinical decision-making and patient counseling. Furthermore, early disease prediction allows diabetes patients and those at risk for diabetes to adopt preventative steps that can postpone the illness's progression and life-threatening consequences. Our research offers several advantages. First, our stacking model predicted diabetes accurately. Using real data to construct prediction models is more realistic and practicable in locations with limited medical resources. Second, a pipeline was constructed to integrate the phases of preprocessing, unbalanced data processing, and a data-driven feature selection technique. Boruta was used to developing significant predictors for detecting the unique classes in the dataset. Hybrid hyperparameter optimization strategies aid in improving the outcome, model building, and model assessment to ensure consistent evaluation of results. Furthermore, because the sample in this study was drawn from a local level, the results may be more representative than those obtained from previous models employing small-scale or small-center data.

However, there are certain restrictions, such as the size of the dataset; large datasets should be included when constructing prediction models, particularly for machine learning algorithms. This does not entail integrating more characteristics but rather features identified as significant in previous models. Despite the fact that this study solely looks at clinical signs. Because this model only works for binary classification, its results will have less impact on multi-class datasets.

In the future, we will address various feature selection approaches and employ more complicated learners to improve the suggested stacking method, such as deep neural networks and a multi-modal approach. Although, in the future, our machine-learning models will require external validation. We validated just the PIMA dataset using a synthetic dataset based on PIMA. As a result, other data sources must be used to validate the models developed in this work.

VIII. CONCLUSION

Diabetes is a silent killer and a chronic condition that can affect many regions of the body. Patients are unable to create enough insulin in their bodies as a result of elevated blood glucose levels. Diabetes prediction can assist both healthcare providers and patients in receiving the correct therapy. We may infer that the proposed approach is the best classification model compared to the other classification models based on assessment parameters such as accuracy and F1-score. Prediction, as well as detection of any disease at the prior level, gives the maximum chance to cure the disease or control its rapid growth and also gives a better way to handle the situation at the earliest viable stage. This research work proposes a prediction model which provides a precise categorization of chronic diseases i.e. diabetes. During Exploratory Data Analysis phase it handle null value, outliers, etc. As the dataset is unbalanced as a consequence, SMOTE data balancing technique was used to balance the dataset. And then applied Boruta; a feature selection technique that gave the best result. Hybrid hyperparameter approach GS-GWO improves the performance of individual models and then finally implementation of stacking classifier enhances the performance of the prediction model. and the proposed model gave 98.84% F1-Score in the case of PIMA, 98% F1-Score after validation of PIMA dataset, 97.3 % F1-Score in case of ADRC and 96.2 % F1-Score in case of FHD dataset. Performance of the proposed Stacking classifier is better in all three dataset. While this prediction indicates that once the patient and his family know about the disease at an early stage they can improve their lifestyle and start medication earliest and can defeat such circumstances otherwise chances can be worst.

ACKNOWLEDGMENT

As a QIP fellow supported by M. J. P. Rohilkhand University, Bareilly, I am doing my Ph.D. from IIT Roorkee, SRE Campus, India. This research is being carried out as part of my doctorate studies. This study and the research effort that went with it would not have been feasible without the extraordinary assistance of my supervisor, Prof. S. C. Sharma, Indian Institute of Technology, Roorkee, SRE Campus, India. The cooperation rendered him in trying out tools, fieldwork, analysis, and interpretation of data and report writing is commendable. Without his support and blessings, the work could not have seen the light of day.

The authors would like to express their sincere gratitude and appreciation to Prof. (Dr.) Prabhat Agarwal, (Diabetologist) S. N. Medical College, Agra, for his invaluable support and contribution to their research paper. Dr. Agarwal generously provided them with the crucial data that served as the foundation for their study. Dr. Prabhat Agarwal's expertise and knowledge in the field of Diabetes were instrumental in ensuring the accuracy and reliability of the data used in their research. His guidance and insightful discussions greatly enhanced their understanding of the subject matter and contributed to the overall quality of this article.

Furthermore, they would like to acknowledge Dr. Sarvesh and Prof. Hemant Yadav, Future Hospital Bareilly for his invaluable support and contribution to their research paper. His expertise, guidance, and data provision have been indispensable, and they are honored to have had the opportunity to work with him.

REFERENCES

[1] K. P. Exarchos, A. Aggelopoulou, A. Oikonomou, T. Biniskou, V. Beli, E. Antoniadou, and K. Kostikas, "Review of artificial intelligence techniques in chronic obstructive lung disease," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 5, pp. 2331–2338, May 2022.

[2] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 156–180, 2021.

[3] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," *PLOS Digit. Health*, vol. 2, no. 1, Jan. 2023, Art. no. e0000082.

[4] J. Gálvez-Goicurla, J. Pagán, A. B. Gago-Veiga, J. M. Moya, and J. L. Ayala, "Cluster-then-classify methodology for the identification of pain episodes in chronic diseases," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 5, pp. 2339–2350, May 2022.

[5] G. Saranya and A. Pravin, "A comprehensive study on disease risk predictions in machine learning," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 10, no. 4, p. 4217, Aug. 2020.

[6] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *J. Amer. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 601–606, Sep. 2011.

[7] A. Budreviciute, S. Damiati, D. K. Sabir, K. Onder, P. Schuller-Goetzburg, G. Plakys, A. Katileviciute, S. Khoja, and R. Kodzius, "Management and prevention strategies for non-communicable diseases (NCDs) and their risk factors," *Frontiers Public Health*, vol. 8, p. 788, Nov. 2020.

[8] C. Wu, T. Zhou, Y. Tian, J. Wu, J. Li, and Z. Liu, "A method for the early prediction of chronic diseases based on short sequential medical data," *Artif. Intell. Med.*, vol. 127, May 2022, Art. no. 102262.

[9] E. M. Senan, M. H. Al-Adhaileh, F. W. Alsaade, T. H. H. Aldhyani, A. A. Alqarni, N. Alsharif, M. I. Uddin, A. H. Alahmadi, M. E. Jadhav, and M. Y. Alzahrani, "Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques," *J. Healthcare Eng.*, vol. 2021, pp. 1–10, Jun. 2021.

[10] P. A. Moreno-Sanchez, "Development and evaluation of an explainable prediction model for chronic kidney disease patients based on ensemble trees," 2021, *arXiv:2105.10368*.

[11] P. S. Baby and T. P. Vital, "Statistical analysis and predicting kidney diseases using machine learning algorithms," *Int. J. Eng. Res.*, vol. V4, no. 7, pp. 206–210, Jul. 2015.

[12] A. T. Mohamed, S. Santhoshkumar, and V. Varadarajan, "Intelligent deep learning based predictive model for coronary heart disease and chronic kidney disease on people with diabetes mellitus," *Malaysian J. Comput. Sci.*, pp. 88–101, Mar. 2022. [Online]. Available: https://ejournal.um.edu.my/index.php/MJCS/article/view/35977

[13] N. Arora, A. Singh, M. Z. N. Al-Dabagh, and S. K. Maitra, "A novel architecture for diabetes patients' prediction using k-means clustering and SVM," *Math. Problems Eng.*, vol. 2022, Aug. 2022, Art. no. 4815521.

[14] *Pima Indians Diabetes Database*. Accessed: Dec. 1, 2022. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[15] E. Tuba, I. Strumberger, T. Bezdan, N. Bacanin, and M. Tuba, "Classification and feature selection method for medical datasets by brain storm optimization algorithm and support vector machine," *Proc. Comput. Sci.*, vol. 162, pp. 307–315, Jan. 2019.

[16] P. A. Moreno-Sanchez, "Chronic kidney disease early diagnosis enhancing by using data mining classification and features selection," in *Proc. Int. Conf. IoT Technol. HealthCare*. Cham, Switzerland: Springer, 2020, pp. 61–76.

[17] S. Hegde and M. R. Mundada, "Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence based feature selection approach," *Int. J. Pervasive Comput. Commun.*, vol. 17, no. 1, pp. 20–36, Feb. 2021.

[18] X. Yang, D. Zhao, F. Yu, A. A. Heidari, Y. Bano, A. Ibrohimov, Y. Liu, Z. Cai, H. Chen, and X. Chen, "An optimized machine learning framework for predicting intradialytic hypotension using indexes of chronic kidney disease-mineral and bone disorders," *Comput. Biol. Med.*, vol. 145, Jun. 2022, Art. no. 105510.

[19] A. Gaber, H. A. Youness, A. Hamdy, H. M. Abdelaal, and A. M. Hassan, "Automatic classification of fatty liver disease based on supervised learning and genetic algorithm," *Appl. Sci.*, vol. 12, no. 1, p. 521, Jan. 2022.

[20] A. S. Abdalrada, "A predictive model for liver disease progression based on logistic regression algorithm," *Periodicals Eng. Natural Sci. (PEN)*, vol. 7, no. 3, pp. 1255–1264, 2019.

[21] A. S. Rahman, F. J. M. Shamrat, Z. Tasnim, J. Roy, and S. A. Hossain, "A comparative study on liver disease prediction using supervised machine learning algorithms," *Int. J. Sci. Technol. Res.*, vol. 8, no. 11, pp. 419–422, 2019.

[22] M. Nilashi, O. B. Ibrahim, H. Ahmadi, and L. Shahmoradi, "An analytical method for diseases prediction using machine learning techniques," *Comput. Chem. Eng.*, vol. 106, pp. 212–223, Nov. 2017.

[23] F. Mostafa, E. Hasan, M. Williamson, and H. Khan, "Statistical machine learning approaches to liver disease prediction," *Livers*, vol. 1, no. 4, pp. 294–312, Dec. 2021.

[24] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health Inf. Sci. Syst.*, vol. 8, no. 1, pp. 1–14, Dec. 2020.

[25] A. Mahabub, "A robust voting approach for diabetes prediction using traditional machine learning techniques," *Social Netw. Appl. Sci.*, vol. 1, no. 12, pp. 1–12, Dec. 2019.

[26] X. Yuan, S. Chen, C. Sun, and L. Yuwen, "A novel early diagnostic framework for chronic diseases with class imbalance," *Sci. Rep.*, vol. 12, no. 1, pp. 1–16, May 2022.

[27] H. Lu and S. Uddin, "A weighted patient network-based framework for predicting chronic diseases using graph neural networks," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, Nov. 2021.

[28] K. Z. Hasan et al., "Performance evaluation of ensemble-based machine learning techniques for prediction of chronic kidney disease," in *Emerging Research in Computing, Information, Communication and Applications*, vol. 1. India: Springer, 2019, pp. 415–426.

[29] T. S. Brisimi, T. Xu, T. Wang, W. Dai, W. G. Adams, and I. C. Paschalidis, "Predicting chronic disease hospitalizations from electronic health records: An interpretable classification approach," *Proc. IEEE*, vol. 106, no. 4, pp. 690–707, Apr. 2018.

[30] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[31] E.-H.-A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informat. Med. Unlocked*, vol. 15, Jan. 2019, Art. no. 100178.

[32] Y. Ren, H. Fei, X. Liang, D. Ji, and M. Cheng, "A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records," *BMC Med. Informat. Decis. Making*, vol. 19, no. S2, pp. 131–138, Apr. 2019.

[33] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *J. Big Data*, vol. 6, no. 1, pp. 1–19, Dec. 2019.

[34] J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, S. Zhu, and Z. Ye, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," *J. Transl. Med.*, vol. 17, no. 1, pp. 1–13, Dec. 2019.

[35] L. H. T. Lam, D. T. Do, D. T. N. Diep, D. L. N. Nguyet, Q. D. Truong, T. T. Tri, H. N. Thanh, and N. Q. K. Le, "Molecular subtype classification of low-grade gliomas using magnetic resonance imaging-based radiomics and machine learning," *NMR Biomed.*, vol. 35, no. 11, Nov. 2022, Art. no. e4792.

[36] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software," *npj Digit. Med.*, vol. 3, no. 1, pp. 1–13, Nov. 2020.

[37] *Gretel Console*. Accessed: Apr. 12, 2023. [Online]. Available: https://console.gretel.ai/use_cases/cards/usecasesynthetic/projects/

[38] X. Zheng, *SMOTE Variants for Imbalanced Binary Classification: Heart 1277 Disease Prediction*, Univ. California, Los Angeles, Los Angeles, CA, USA, 2020.

[39] J. Yang and J. Guan, "A heart disease prediction model based on feature optimization and smote-Xgboost algorithm," *Information*, vol. 13, no. 10, p. 475, Oct. 2022.

[40] A. Kishor and C. Chakraborty, "Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE," *Int. J. Syst. Assurance Eng. Manage.*, pp. 1–9, Jun. 2021.

[41] P. Chittora, S. Chaurasia, P. Chakrabarti, G. Kumawat, T. Chakrabarti, Z. Leonowicz, M. Jasinski, L. Jasinski, R. Gono, E. Jasinska, and V. Bolshev, "Prediction of chronic kidney disease—A machine learning perspective," *IEEE Access*, vol. 9, pp. 17312–17334, 2021.

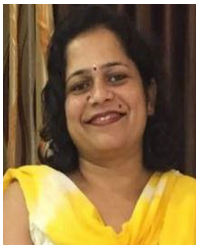
[42] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egyptian Informat. J.*, vol. 19, no. 3, pp. 179–189, Nov. 2018.

[43] H. Polat, H. D. Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," *J. Med. Syst.*, vol. 41, no. 4, pp. 1–11, Apr. 2017.

[44] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," *J. Big Data*, vol. 2, no. 1, pp. 1–36, Dec. 2015.

[45] K. Tadist, S. Najah, N. S. Nikolov, F. Mrabti, and A. Zahi, "Feature selection methods and genomic big data: A systematic review," *J. Big Data*, vol. 6, no. 1, pp. 1–24, Dec. 2019.

[46] R. R. Rajalaxmi, "A hybrid binary cuckoo search and genetic algorithm for feature selection in type-2 diabetes," *Current Bioinf.*, vol. 11, no. 4, pp. 490–499, Aug. 2016.
[47] N. Kushmerick, *Wrapper Induction for Information Extraction*, Univ. Washington, Seattle, WA, USA, 1997.
[48] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010.
[49] M. W. Mwadulo, "A review on feature selection methods for classification tasks," *Int. J. Comput. Appl. Technol. Res.*, vol. 5, no. 6, pp. 395–402, 2016.
[50] J. Abdollahi and B. Nouri-Moghaddam, "Feature selection for medical diagnosis: Evaluation for using a hybrid stacked-genetic approach in the diagnosis of heart disease," 2021, *arXiv:2103.08175*.
[51] F. F. Firdaus, H. A. Nugroho, and I. Soesanti, "Deep neural network with hyperparameter tuning for detection of heart disease," in *Proc. IEEE Asia Pacific Conf. Wireless Mobile (APWiMob)*, Apr. 2021, pp. 59–65.
[52] R. Valarmathi and T. Sheela, "Heart disease prediction using hyper parameter optimization (HPO) tuning," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 103033.
[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2012.
[54] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014.
[55] R. Mahadeva, M. Kumar, S. P. Patole, and G. Manik, "Desalination plant performance prediction model using grey wolf optimizer based ANN approach," *IEEE Access*, vol. 10, pp. 34550–34561, 2022.
[56] R. Mahadeva, M. Kumar, S. P. Patole, and G. Manik, "PID control design using AGPSO technique and its application in TITO reverse osmosis desalination plant," *IEEE Access*, vol. 10, pp. 125881–125892, 2022.
[57] N. M. Sallam, A. I. Saleh, H. Arafat Ali, and M. M. Abdelsalam, "An efficient strategy for blood diseases detection based on grey wolf optimization as feature selection and machine learning techniques," *Appl. Sci.*, vol. 12, no. 21, Oct. 2022, Art. no. 10760.
[58] A. I. Naimi and L. B. Balzer, "Stacked generalization: An introduction to super learning," *Eur. J. Epidemiol.*, vol. 33, no. 5, pp. 459–464, May 2018.
[59] M. Kumar, S. Singhal, S. Shekhar, B. Sharma, and G. Srivastava, "Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning," *Sustainability*, vol. 14, no. 21, Oct. 2022, Art. no. 13998.
[60] R. Mahadeva, M. Kumar, S. P. Patole, and G. Manik, "An optimized PSO-ANN model for improved prediction of water treatment desalination plant performance," *Water Supply*, vol. 22, no. 3, pp. 2874–2882, Mar. 2022.
[61] R. Mahadeva, M. Kumar, S. P. Patole, and G. Manik, "Employing artificial neural network for accurate modeling, simulation and performance analysis of an RO-based desalination process," *Sustain. Comput., Informat. Syst.*, vol. 35, Sep. 2022, Art. no. 100735.
[62] R. Mahadeva, R. Mehta, G. Manik, and A. Bhattacharya, "An experimental and computational investigation of poly(piperizinamide) thin film composite membrane for salts separation from water using artificial neural network," *Desalination Water Treatment*, vol. 224, pp. 106–121, Jan. 2021.
[63] R. Mahadeva, M. Kumar, G. Manik, and S. P. Patole, "Modeling, simulation, and optimization of the membrane performance of seawater reverse osmosis desalination plant using neural network and fuzzy based soft computing techniques," *Desalination Water Treatment*, vol. 229, pp. 17–30, Jan. 2021.
[64] R. Mahadeva, "Modelling and simulation of reverse osmosis system using PSO-ANN prediction technique," in *Soft Computing: Theories and Applications*. Cham, Switzerland: Springer, 2020, pp. 1209–1219.
[65] S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, and M. H. Rahman, "Prevalence and early prediction of diabetes using machine learning in north kashmir: A case study of district bandipora," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, Oct. 2022.
[66] P. Madan, V. Singh, V. Chaudhari, Y. Albagory, A. Dumka, R. Singh, A. Gehlot, M. Rashid, S. S. Alshamrani, and A. S. AlGhamdi, "An optimization-based diabetes prediction model using CNN and bi-directional LSTM in real-time environment," *Appl. Sci.*, vol. 12, no. 8, p. 3989, Apr. 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/8/3989
[67] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, "Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques," *Mobile Inf. Syst.*, vol. 2022, pp. 1–16, Mar. 2022.
[68] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, Jan. 1992.
[69] S.-A. N. Alexandropoulos, C. K. Aridas, S. B. Kotsiantis, and M. N. Vrahatis, "Stacking strong ensembles of classifiers," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.* Cham, Switzerland: Springer, 2019, pp. 545–556.
[70] R. Liu, Y. Zhan, X. Liu, Y. Zhang, L. Gui, Y. Qu, H. Nan, and Y. Jiang, "Stacking ensemble method for gestational diabetes mellitus prediction in Chinese pregnant women: A prospective cohort study," *J. Healthcare Eng.*, vol. 2022, pp. 1–14, Sep. 2022.
[71] T. Shen, H. Yu, and Y.-Z. Wang, "Discrimination of gentiana and its related species using IR spectroscopy combined with feature selection and stacked generalization," *Molecules*, vol. 25, no. 6, p. 1442, Mar. 2020.
[72] H. Salah and S. Srinivas, "Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents," *Sci. Rep.*, vol. 12, no. 1, Dec. 2022, Art. no. 21905.
[73] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, pp. 1–9, Dec. 2020.
[74] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan, "Explainable AI (XAI): Core ideas, techniques, and solutions," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–33, Sep. 2023.
[75] T. H. Vo, N. T. K. Nguyen, Q. H. Kha, and N. Q. K. Le, "On the road to explainable AI in drug-drug interactions prediction: A systematic review," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 2112–2123, Jan. 2022.
[76] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020.
[77] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, p. 3317, Mar. 2021.
[78] I. Rodríguez-Rodríguez, J.-V. Rodríguez, W. L. Woo, B. Wei, and D.-J. Pardo-Quiles, "A comparison of feature selection and forecasting machine learning algorithms for predicting glycaemia in type 1 diabetes mellitus," *Appl. Sci.*, vol. 11, no. 4, p. 1742, Feb. 2021.
[79] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput. Appl.*, vol. 35, pp. 16157–16173, Mar. 2022.
[80] S. A. Alex, N. Jhanjhi, M. Humayun, A. O. Ibrahim, and A. W. Abulfaraj, "Deep LSTM model for diabetes prediction with class balancing by SMOTE," *Electronics*, vol. 11, no. 17, p. 2737, Aug. 2022.
[81] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, Apr. 2022.



POOJA YADAV (Member, IEEE) received the B.Tech. degree in computer science and engineering from Purvanchal University, India, in 2002, and the M.Tech. degree in computer science and engineering from Aligarh Muslim University, Aligarh, India, in 2005. She is currently pursuing the Ph.D. degree with the Indian Institute of Technology Roorkee, India. She is an Assistant Professor with the Department of Computer Science and Information Technology, Faculty of Engineering and Technology, M. J. P. Rohilkhand University, Bareilly, Uttar Pradesh, India. She has 15 years of experience in academics. She is the author of four books and some book chapters. She has published research papers in various journals and conferences of international repute. Her current research interests include the IoT and machine learning in the healthcare sector. She is a member of the Institute of Engineers.

S. C. SHARMA received the M.Sc. (Elect.), M.Tech. (Elect. and Comm. Engg.), and Ph.D. (Elect. and Comp Engg.) degrees from IIT Roorkee (erstwhile University of Roorkee), in 1981, 1983, and 1992, respectively. In 1983, he started his career as a Research and Development Engineer then joined the teaching profession in January 1984 with IIT Roorkee and continues to date. He has more than 35 years of teaching and research experience at IIT Roorkee. He has published over 300 research papers in national and international journals (152)/conferences (150) and supervised more than 30 projects/dissertations of PG students. He has supervised 20 Ph.D. students in the area of computer networking, wireless networks, computer communication, cloud and its security, and mobile computing and continues supervising Ph.D. students in the same area. He has successfully completed several major research projects funded by various Government Agencies, such as AICTE, CSIR, UGC, MHRD, DST, DRDO, and many minor research projects related to communication and SAW filter design sponsored by the Government of India. He was a Research Scientist with FMH, Munich, Germany, and visited many countries (U.K., France, Germany, Italy, Switzerland, Canada, United Arab Emirates, Thailand, and The Netherlands) related to research work. He has chaired sessions at international conferences and delivered invited talks at various forums. He is an active reviewer of the IEEE journals and an editor of various reputed international and national journals. He is an Honorary Member of NSBE, ISOC, and IAENG, USA. He was also the Group Leader of the Electronics and Instrumentation Engineering Department, BITS-Pilani-Dubai Campus, from August 2003 to August 2005. He is currently continuing as a Professor with IIT Roorkee, Saharanpur Campus. IIT Roorkee has awarded him the Khosla Annual Research Prize for the Best Research Paper. His many research papers have been awarded by national and international committees and journals.

RAJESH MAHADEVA (Member, IEEE) received the B.E. degree in electronics and instrumentation engineering from the Samrat Ashok Technological Institute (SATI), Vidisha, Madhya Pradesh, India, in 2006, the M.Tech. degree in control and instrumentation engineering from the National Institute of Technology (NIT), Jalandhar, Punjab, India, in 2009, and the Ph.D. degree from the Department of Polymer and Process Engineering, Indian Institute of Technology (IIT) Roorkee, Uttarakhand, India, in 2022. From 2011 to 2017, he was an Assistant Professor with the Technocrats Institute of Technology (TIT), Bhopal, and Marwadi University (MU), Rajkot, India. He is currently a Researcher with the Department of Physics, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates. His research interests include the modeling, simulation, optimization, and control of desalination and water treatment plants/processes using artificial intelligence techniques.

SHASHIKANT P. PATOLE (Member, IEEE) received the B.Sc., M.Sc., and M.Phil. degrees in physics from the University of Pune (UoP), India, and the Ph.D. degree in nanoscience and technology from Sungkyunkwan University (SKKU), Suwon, South Korea, in 2010. Since 2017, he has been an Assistant Professor with the Physics Department, Khalifa University of Science and Technology (KU), Abu Dhabi, United Arab Emirates. Before joining KU, he was an Entrepreneur and a Postdoctoral Researcher with the King Abdulla University of Science and Technology (KAUST), Saudi Arabia. In KAUST, he was a Founding Member of the Laboratory for Carbon Nanostructures and the Co-Founder of Graphene Crystal startup company. He has published more than 50 articles. His research interests include the development and commercialization of advanced quantum materials for sustainable energy and environment, carbon nanotubes, graphene, and other 2D materials in membrane technology, structural composites and energy, optoelectronics, electron field emission, photovoltaic, and aberration-corrected transmission electron microscopy.