IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 October 2023, accepted 13 November 2023, date of publication 28 November 2023,
date of current version 1 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3335314

RESEARCH ARTICLE

Particle Swarm Optimization-Based Random Forest Framework for the Classification of Chronic Diseases

AKANSHA SINGH[1], NUPUR PRAKASH[2], AND ANURAG JAIN[1]
[1]University School of Information, Communication, and Technology, Guru Gobind Singh Indraprastha University, Delhi 110078, India
[2]Department of Computer Science and Engineering, The Northcap University, Gurugram 122017, India

Corresponding author: Anurag Jain (anurag@ipu.ac.in)

ABSTRACT In this paper, a hybrid metaheuristic-based Machine learning approach has been propounded for the classification of various Chronic Diseases (CDs). The CDs often get misdiagnosed due to various issues viz., similar and overlapping symptoms, sensitive devices, lack of clinical experts, etc. Based on the above issues, this study has utilized a fusion of Particle Swarm Optimization with Random Forest (PSORF) for the automatic identification of CDs. The approach PSORF comprises of two main components: PSO for obtaining the minimal optimal feature set, also to optimize the performance of the RF classifier, and RF classifier for the classification of multiple CDs. In this research, five different CD datasets have been deployed onto a series of experiments have been conducted to identify the best approach for the classification of CDs. To address the issues of imbalanced and incomplete data in the datasets used, Synthetic Minority Oversampling Technique (SMOTE) and Expected Minimization (EM) Imputation techniques have been applied before training the model. This ensures the data quality is improved before being used for analysis. Furthermore, the performance of the PSO and RF classifiers has been compared with other metaheuristic and ML classifiers in terms of different performance metrics. For this purpose, Friedman's tests have been employed to calculate the mean ranks of all the classifiers across all the datasets for different metrics. The results showed that the proposed technique achieved the highest mean rank in terms of Accuracy, F-measure, and Receiver Operating Characteristics (ROC) across all five datasets.

INDEX TERMS Chronic diseases, machine learning, metaheuristic techniques, multi-classification, PSO, SMOTE.

I. INTRODUCTION

Chronic diseases (CD) are long-lasting diseases causing millions of deaths and disability worldwide. Especially, post-pandemic CDs are on the rise as the virus not only affects the lungs but also the other parts of the body.[1] Such diseases cannot be cured completely but can be controlled and treated only if detected early.[2] In regard to the classification of CDs, this study has focused on three different domains of CDs as shown in Figure 1.

CDs such as heart disease, lung disease, cancer, diabetes, etc., are the leading cause of death and disability worldwide. These are such diseases whose symptoms show up at the later stages which makes it even harder to treat them. The most prevalent form of heart disease is Coronary Artery Disease (CAD), which occurs when a major artery (such as the Left Anterior Descending (LAD), Left Circumference Artery (LCA), or Right Coronary Artery (RCX)) becomes narrowed due to stenosis. The deaths resulting from CAD were reported as 382,820 in 2020 [1]. The symptom of CAD includes shortness of breath, chest pain, chest tightness, and sweats.

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

[1] Post-COVID symptoms and effects, accessed on 19/06/2023.

[2] Early diagnosis of Chronic diseases, accessed on 19/06/23.

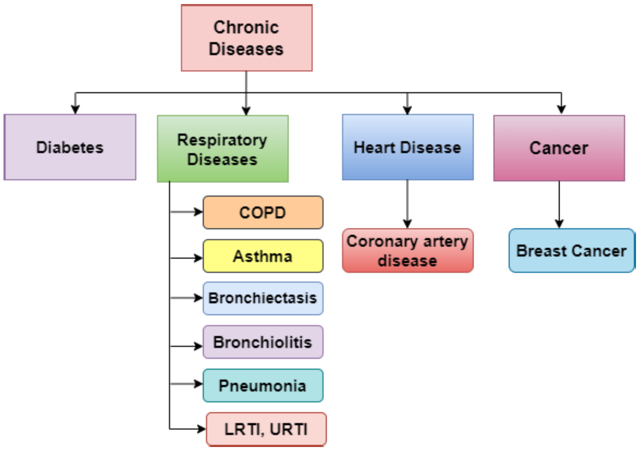


FIGURE 1. Taxonomy of the chronic diseases utilized in this study.

The next group of CDs is respiratory diseases consisting of Chronic Obstructive Pulmonary Disease (COPD), Asthma, Pneumonia, Tuberculosis (TB), etc [2]. These are the group of diseases that has been greatly affected by the COVID-19 pandemic as it has directly attacked the lungs making them less immune to other respiratory diseases. The number of deaths reported due to COPD, Asthma, Lung cancer, and Pneumonia are 3.2, 260, 1.8, and 2.4 million respectively.[3] Such chronic diseases give rise to a range of symptoms, such as shortness of breath, excessive mucus production, chest pain, tightness in the chest, coughing, and many more. Breast cancer is currently the most prevalent form of cancer, resulting in 685,000 fatalities worldwide.[4] Similarly, Diabetes, a chronic metabolic disease contributes to 1.5 million deaths each year.[5] It is riskier as it can affect the other major organs such as the Heart and kidneys. Although there are various fundamentally well-organized primary care approaches for treating CDs such as Spirometry pulmonary functional test for COPD, X-rays, scans for other lung diseases, surgical removal, radiation therapy, mammograms for Breast cancer [3], and Angiography for CAD, there are various issues related to such treatments mentioned as follows:

- Overdiagnosis in case of mammograms, radiation injury during chemotherapy.[6]
- The miniature size of tumors and lung nodules cannot be read clearly by clinical experts.[7]
- Misclassification of diseases due to similar and overlapping symptoms.[8]
- Expensive medical procedures such as Angiography.[9]
- Rely on manual labeling by doctors which is time-consuming and laborious.[10]

Therefore, in response to the aforementioned issues, several researchers have proposed implementing computer-aided diagnosis, which can simplify the work of physicians and decrease the likelihood of misdiagnosis. Previous research on various CDs [4], [5], [6], [7], [8], [9], [10], [11] has demonstrated that it is feasible to detect and categorize CDs using Machine Learning (ML) technology. However, such approaches have shown lower performances and exhibit some limitations such as imbalanced data, accuracy paradox, missing data problems, slow convergence by metaheuristic techniques like Genetic Algorithm (GA), etc. Some studies even achieved excellent performances. Howbeit, that might be the case of accuracy paradox, a condition that occurs when the models achieve excellent performance by training on biased or imbalanced data. Hence, this study aims at providing a metaheuristic-based framework Particle Swarm Optimization based Random Forest (PSORF) that can diagnose CRDs efficiently while dealing with issues found in previous studies. In this study, the problem of imbalanced data and missing data has been rectified by utilizing the SMOTE filter and EM Imputation method respectively. The slow convergence problem of GA has been resolved by using PSO. The ability of PSO to search larger spaces efficiently, being less computationally expensive, and faster convergence has made it an effective and efficient global search technique as compared to other techniques such as the Genetic, Bat, and Firefly algorithms. Similarly, Random Forest (RF) has a great advantage over other ML techniques such as its ability to deal with missing and imbalanced data, reduce overfitting by using an ensemble of various decision trees, etc. The performance of the proposed approach has been evaluated through a series of experiments on five different chronic disease datasets and compared to other benchmark metaheuristics and ML techniques. The remarkable performance of the proposed approach is evident from the results, surpassing other techniques. The major contributions of this study have been listed as follows-

- A hybrid metaheuristic-based ML classifier (PSORF) has been proposed that can not only diagnose a disease but can also differentiate various similar CDs based on symptomatic information.
- EM Imputation and SMOTE techniques have been employed to fill in the missing values and treat imbalance data problems respectively.
- Performance of different metaheuristic techniques such as PSO, GA, Bat, and Firefly Algorithm (FA) has been compared using radar charts.
- Friedman's Test has been utilized to corroborate the performance of the proposed approach with the other ML classifiers by comparing their mean ranks.

The remaining sections of this paper are structured as follows: Section II discusses the previous research and

[3] Respiratory disease, Number of deaths reported, accessed on 19/06/23.
[4] Breast cancer, factsheet for Breast Cancer", accessed on 20/06/23.
[5] Diabetes", accessed on 20/09/23.
[6] Overdiagnosis of mammograms, accessed on 22/06/23.
[7] Missed detection of lung cancer, accessed on 22/06/23.
[8] Misdiagnosis of lung diseases, accessed on 22/06/23.
[9] Coronary artery Angiogram, accessed on 22/06/23.
[10] Mannual data labeling, accessed on 22/06/23.

identifies gaps in the use of ML techniques for detecting CDs. Section III outlines the materials and methods utilized in this study. In addition, it explains the benchmarks feature selection and ML techniques in brief. Furthermore, it explains the proposed methodology and all its stages in detail. Section IV illustrates the experimental work carried out on different datasets using the proposed approach and shows the comparison of the proposed approach with other ML and metaheuristic techniques. It further explains the limitations and future work of the study. Section V concludes the study.

II. RELATED WORK AND RESEARCH GAPS

In the literature, several researchers have examined numerous ML models for the detection of various CDs to help clinical decision-making. In this regard, this section discusses multiple works done for the classification of Breast Cancer, heart, Diabetes, and respiratory diseases as shown in Table 1, 2, 3, and 4 and also identifies the research gaps.

Upon review of prior research, it was found that many studies employed metaheuristic optimization algorithms for feature selection and different machine learning (ML) and deep learning (DL) models for disease classification, as outlined in Tables 1, 2, 3, and 4. While these previous studies have yielded promising outcomes, there are still some areas for further research and improvement, as described below.

- Imbalance dataset: It must be acknowledged that previous research has frequently depended on imbalanced datasets to predict diseases, producing biased outcomes. However, a study conducted by Zhang et al. [21] resolved this issue by utilizing the SMOTE filter. It is crucial to meticulously scrutinize potential biases when interpreting research findings.
- Missing data: In this study, it was found that the Exasens dataset contains some missing values that must be addressed before being used in the training model. If left untreated, such values can significantly impact the accuracy of the classification model. Previous studies by Ramachandra and Murthy [27] and Gill and Pathwar [28] did not address these missing values. However, Amutha and Sekar [26] utilized the KNN Imputation method to address this issue. While this method is effective, adaptive, and flexible, it can be susceptible to outliers and is computationally expensive.
- Lower performances: Previous studies have clearly demonstrated that certain datasets exhibit lower performance levels due to missing data, lack of feature selection, and high computational models [22], [29], [30]. It has been observed that studies that employed metaheuristic-based ML classifiers outperformed those using DL models when comparing studies that utilized the same dataset. While DL models are known for their automatic feature selection, it is important to note that tuning these features and the model's parameters can consume a significant amount of computational resources. On the other hand, utilizing metaheuristic optimization algorithms for feature selection with ML classifiers greatly reduces the computational power required. Therefore, it can be concluded that Metaheuristic Optimization (MHO) based ML classifiers have been shown to outperform DL models.
- Accuracy Paradox: Despite various issues and research gaps, previous studies achieved excellent performances. The accuracy paradox may be at play here. Even though the training model achieves high accuracy levels, it has low predictive value. This is especially true when handling an imbalanced Breast cancer dataset, where the accuracy rate can be over 97% in all cases [12], [13], [14], [15], [16], [17]. However, such a model trained on this data may not perform well in identifying cancer patients in real-life situations, despite producing accurate training results due to a high proportion of cancer patients' examples.
- No statistical testing: It's worth noting that only a few studies have been found in the literature that utilized statistical testing to validate their models and achieve optimal performance [13] and [22]. Most studies instead compared various ML and DL models using different performance metrics to determine the top performer. However, these results were not adequately explained in those studies.

In order to create a reliable and effective model, this study has addressed all of the research gaps mentioned previously. The issue of imbalanced and missing data was tackled in section III, while section IV thoroughly explains and confirms the classification performance of the proposed model.

III. MATERIALS AND METHODS

In this section, the materials and methods utilized in this study have been examined. It describes the different datasets employed in this study and then discusses the benchmark metaheuristic and ML classification techniques. It further showcases the different stages of the proposed methodology in detail.

A. DATASETS

This study has employed five publicly available datasets as evaluation benchmarks: the International Conference on Biomedical Health Informatics (ICBHI) lung sound database [35], Wisconsin Breast Cancer Dataset (WBCD) [36], Z-Alizadehsani dataset [37], Exasens dataset [38], and Diabetes dataset [39] collected from UCI library, Kaggle, and dataworld. For ease purpose, datasets ICBHI, WBCD, Z-Alizadehsani, Exasens, and Diabetes have been specified as D1, D2, D3, D4, and D5 respectively. Detailed information regarding each dataset has been presented in Table 5.

In this study structured data consisting of symptomatic information in accordance with the respective diseases has been considered for the evaluation of ML classifiers for classifying Chronic diseases. The distribution of instances

TABLE 1. Previous works done for the detection of breast cancer using different feature selection and ML approaches on the wisconsin breast cancer dataset.

Author	Year	Feature Selection	Classifiers	Accuracy	Limitations
Oladele, T. O. et.al., [12]	2021	Ant Colony Optimization (ACO), PSO	Support vector Machine (SVM), C4.5, Naive Bayes (NB), Convolution neural network (CNN), Neural Network (NN), Logistic Regression (LR), RF	97.13% for both ACO and PSO	Imbalance data problem, undecidability between ACO and PSO, no statistical tests were performed
Ogundokun, R. O. et. al., [13]	2022	PSO	ANN, CNN, SVM, Grid Search	98.5% (CNN) 99.2%	Imbalance data, High computational deep learning model
Sahu, B. et.al., [14]	2019	Principle Component Analysis (PCA)	ANN, SVM, K-Nearest Neighbor (KNN), RF	97% (PCA+ANN)	Lower performance, Imbalance data, no statistical tests were performed
Olorunsola, B. J. [15]	2021	GA, PSO, Harmony search, Tabu search	SVM, C4.5, NB, KNN, NN, LR, RF	97.1% (PSO+RF)	Lower performance, Center bias operator problem in Harmony Searchy, Imbalance data, no statistical tests were performed
Guo, Z. [16]	2022	GA, PSO, Open Source Development Model Algorithm (ODMA)	Multilayer Perceptron (MLP)	98.79%	Imbalance data, no statistical tests were performed
Jia, X. [17]	2022	Wolf Optimization Algorithm (WOA)	SVM	99.02%	Imbalance data, Center bias operator problem in WOA, no statistical tests were performed
Huang, H. [18]	2019	Firefly Optimization Algorithm (FOA)	SVM	93.83%	Lower performance, Center bias operator problem in FOA, Imbalance data, no statistical tests were performed

TABLE 2. Previous works done for the detection of Coronary artery disease using different feature selection and ML approaches on the Z-alizadehsani dataset.

Author	Year	Feature Selection	Classifiers	Accuracy	Limitations
Gupta, A. et. al., [19]	2021	PSO, GA, FA, Bat, Gravitational, Dragonfly, FAMD	Ensemble (RF+Extra tree)	97.37%	Imbalance data, Center bias operator problem in FA, lower performance
Fajri, Y. A., [20]	2022	Bee Swarm and Q-learning	SVM, RF, LightGBM, XGBoost	90.1%	Lower performance, increased time complexity, no statistical tests were performed
Zhang, S. et. al., [21]	2022	SMOTE	LightGBM	94.7%	Lower performance, no feature selection, no statistical tests were performed
Hassannataj Joloudari, et.al., [22]	2022	Genetic algorithm	SVM+ANOVA test	89.45%	lower performance, Imbalance data
Kolukisa, B., et.al., [23]	2023	Computational FS methods	MLP	91.78%	Lower performance, Imbalance data, statistical tests were not performed
Kolukisa, B., et.al., [24]	2020	No feature selection	KNN, SVM, LR, Linear Discriminant Analysis (LDA), NB, Ensemble	88.38%	Lower performance, Imbalance data, statistical tests were not performed, no feature selection
Singh, A., et.al., [25]	2021	InfoGain, GainRatio	RF, NB, MLP	95.70% for RF	Lower performance, Imbalance data, statistical tests were not performed

TABLE 3. Previous works done for the detection of Diabetes using different feature selection and ML approaches on the vanderbilt diabetes dataset.

Author	Year	Feature Selection	Classifiers	Accuracy	Limitations
Amutha, S., et. al., [26]	2023	Binarised Grey Wolf (GW) and Whale optimization (WO) technique	Grid search based SVM	98.71%	Center bias operator problem in WOA and GWOA, missing data, no statistical tests were performed
AC, R., et. al., [27]	2023	No feature selection	Logistic Regression, stacking	93%	No feature selection, Lower performance, missing data, no statistical tests
Gill, S., et.al., [28]	2022	GA	RF	93.95%	Lower performance, missing data, no statistical tests were performed
Rajendra, P., et.al., [29]	2021	Chi-square	Ensemble technique	93%	Lower performance, missing data

into different classes corresponding to different diseases is shown in Figure 2.

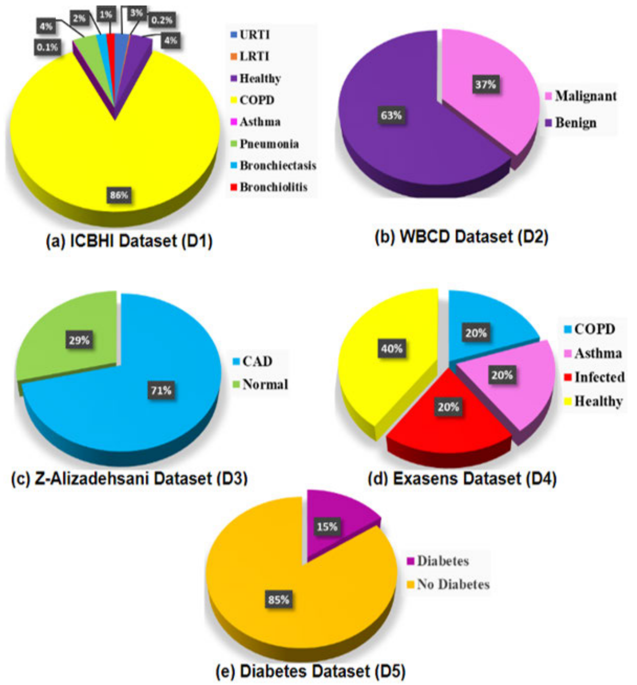
Further details regarding the datasets are mentioned as follows:

TABLE 4. Previous works done for the detection of Respiratory Disease using different feature selection and ML approaches on the ICBHI and exasens dataset.

Author	Year	Feature Selection	Classifiers	Accuracy	Limitations
Dhar, J. [30]	2021	GA	Ensemble model, GB, NB, RF, Support Vector Classifier (SVC), KNN, LR	98.2%	Imbalance data problem, no statistical tests were performed
Irshad, R. R., et. al., [31]	2023	Tasmanian Devil optimization (TDO)	DCNN	99.1%	Imbalance data, High computational deep learning model, missing data
Zarrin, P. S., et.al., [32]	2020	Grey Wolf Optimization (GWO)	XGBoost, SVM, Gaussian Naive Bayes (GNB), LR, ANN	91.25%	Lower performance, Imbalance data, missingness problem, no statistical tests were performed
Petmezas, G. [33]	2022	Short-time Fourier Transform (STFT)	Large Short term Memory (LSTM)	76.39%	Lower performance, Imbalance data, no statistical tests were performed
Ali, S. W. [34]	2023	Mel-Frequency cepstral coefficient (MFCC)	MusicANN, VGGish, and OpenL3	81%	Imbalance data, no statistical tests were performed

TABLE 5. Deployement of datasets for the identification of chronic diseases.

Description	ICBHI	WBCD	Z-Alizadehsani	Exasens	Diabetes
Data Size	10120 * 8	569 * 32	303 * 59	399 * 9	390 * 16
Data type	Sound	Tabular data	Tabular data	Tabular data	Tabular data
Missing values	No	No	No	Yes	No
Imbalance dataset	Yes	Yes	Yes	Yes	Yes
Disease	COPD, Asthma, BE, Bronchiolitis, LRTI, URTI, Pneumonia	Breast Cancer	Coronary Artery Disease	COPD, Asthma	Diabetes
Classification	Multi-classification	Binary	Binary	Binary	Binary
Target class	Diseases	Diagnosis	Cath	Diagnosis	Diabetes



(a) ICBHI Dataset (D1)

(b) WBCD Dataset (D2)

(c) Z-Alizadehsani Dataset (D3)

(d) Exasens Dataset (D4)

(e) Diabetes Dataset (D5)

FIGURE 2. Distribution of instances into the number of target classes corresponding to datasets a) ICBHI dataset D1, b) WBCD dataset D2, c) Z-Alizadehsani dataset D3, d) Exasens dataset D4, e) Diabetes dataset D5.

- ICBHI Respiratory Sound Database: The dataset was collected by two research teams in Portugal and Greece. It consists of 920 annotated recordings of length ranging from 10s to 90s making it a total of 5.5 hours of recordings. The recordings are collected from 126 patients. It contains 6898 respiratory cycles wherein 1864, 886, and 506 contain crackles, wheezes and both crackles and wheeze respectively [40], [41].
- WBCD: The dataset was created at the University of Wisconsin Hospitals in 1992. The attribute "diagnosis" has been denoted as the class label that classifies the tumor as Malignant (M) and Benign (B). In the literature, the majority of the papers worked on unstructured data for Breast cancer like mammograms [42], [43].
- Z-Alizadehsani Dataset: The data was collected from heart disease patients at Shaheed Rajaei Cardiovascular, Medical, and Research Center, Tehran, Iran. This dataset is an extension of the Z-Alizadehsani dataset and was collected from the UCI library. In this dataset, the information about the major three arteries has been added increasing the total number of attributes to 59. The attributes are grouped into four categories: demographic information, symptoms and examination, ECG, and laboratory and echo features [40], [44].
- Exasens Dataset: The dataset was collected at Research Center Borstel, Germany. It contains information regarding the four groups of saliva samples namely, COPD, Asthma, Infected, and healthy [40].
- Diabetes Dataset: The dataset utilized in this study is a modified version of the original Vanderbilt Diabetes dataset [45] originated from a study conducted on rural African Americans. The original dataset consisted of

patients with several missing values. Before deploying the dataset into this study, 13 patients with heavily missing data were excluded.[11]

B. BENCHMARK TECHNIQUES

This section discusses the benchmark techniques utilized in this study for comparing and validating the performance of the proposed approach. As mentioned earlier, the proposed approach comprises two components i.e., PSO and RF. Hence, for comparison purposes, two sets of benchmark techniques have been utilized. One set is for comparing feature selection techniques and another set is for comparing proposed approaches with state-of-art classifiers.

1) FEATURE SELECTION

In this study, to compare and validate the performance of PSO, three benchmark metaheuristic optimization techniques GA [15], [16], [19], Bat [19], and FA [19] have been employed. These algorithms are population-based algorithms where the agents perform both local and global searches. They are iterative in nature. They generally start from a randomly chosen solution and move forward. The goal is to find an optimal solution at each iteration until no further improvements can be made. Also, It is not advisable to use the Firefly algorithm as one of the benchmark techniques due to its ''center bias operator'' problem [46] because this operator enables the algorithm to optimize its function in a way that places its respective optima in the center of the feasible set. Despite this, numerous studies in the literature have utilized this algorithm for feature selection and tuning of hyper-parameters of ML classifiers. For comparison purposes, this study has incorporated both types of MHO algorithms, one with and others without a center bias operator problem.

2) ML CLASSIFIERS

This section discusses the cutting-edge classifiers that were employed to assess and verify the effectiveness of the proposed method.

- Naïve Bayes: This supervised learning classifier is an amalgamation of two terms: The term ''naive'' indicates that the algorithm assumes conditional independence between all features, given the value of the class variable. On the other hand, the term ''Bayes'' indicates that the method is based on the Bayes theorem [12], [27]. This theorem describes the relationship between the class variable (denoted as z) and the dependent feature vectors (y_1 through y_n). as shown in (1).

$$P(z|y_1,\ldots..y_n) = \frac{(P(z)P(y_1\ldots..y_n|z))}{P(y_1\ldots..y_n)} \tag{1}$$

  There are different versions of Naïve Bayes which differ only in terms of the assumption they make regarding the distribution $P(y_i|z)$ [32].
- Multilayer Perceptron (MLP): It is the simplest form of neural network that learns a function $f(\cdot) : \mathbb{R}^p \longrightarrow \mathbb{R}^q$ by training on a dataset where p is the number of dimensions for the input and q is the number of dimensions for the output [13], [16]. The model consists of three layers: the ''Input layer'' consists of a set of neurons $x_i|x_1, x_2, \ldots..x_n$ indicating the input features, the middle layer is the ''Hidden layer'' consisting of one or more layers containing neurons that transform the previous layer values into a weighted linear summation and then apply a non-linear activation function $g(\cdot) : \mathbb{R}^p \longrightarrow \mathbb{R}^q$, and the last layer ''Output layer'' that receives the input from the hidden layer and transform it into the output values [25].
- Sequential Minimal Optimization (SMO): A supervised learning algorithm designed for the training of SVM as its training requires solving large complicated Quadratic Programming (QP) optimization problems. This problem becomes more cumbersome when dealing with large datasets leading to a running time of $O(N^3)$ [47]. SMO breaks these large QP problems into small QP problems which then can be solved analytically. All these calculations make SMO scale between linear or quadratic in the training set size hence making it faster than SVM.
- Bagging. It is an averaging ensemble classifier that builds several estimators independently and then averages their predictors. The idea is that the combined estimators perform better than single estimators due to the reduction in variance. It works best with strong and complex models as they reduce overfitting [23].

C. PROPOSED METHODOLOGY

This section introduces the details of the proposed approach PSO-RF for the multiclassification of Chronic Diseases. Additionally, various stages of the proposed approach have been exhibited in Figure 3.

The key elements of each stage are briefly elaborated on as follows:

1) STAGE 1: DATA PREPROCESSING

In this stage, the original raw data has been treated in terms of quantity and quality by having it pass through different sub-stages to enhance the performance of the proposed approach. The various sub-stages are shown in Figure 4.

The datasets were first checked for their types. Among all the datasets, dataset D1 was unstructured and needed to be converted into structured data using Python programming. Hence, the .csv file containing the patient id and disease has been aligned with the .txt file of different .wav files to get a structured file. In addition, it has been observed from Table 5 that the Exasens dataset suffers from a missingness problem. The dataset consists of 33.36% of the whole data missing values. In this regard, this study has deployed

[11] Diabetes dataset, Modified dataset by Robert Hoyt'', accessed on 20/09/23.

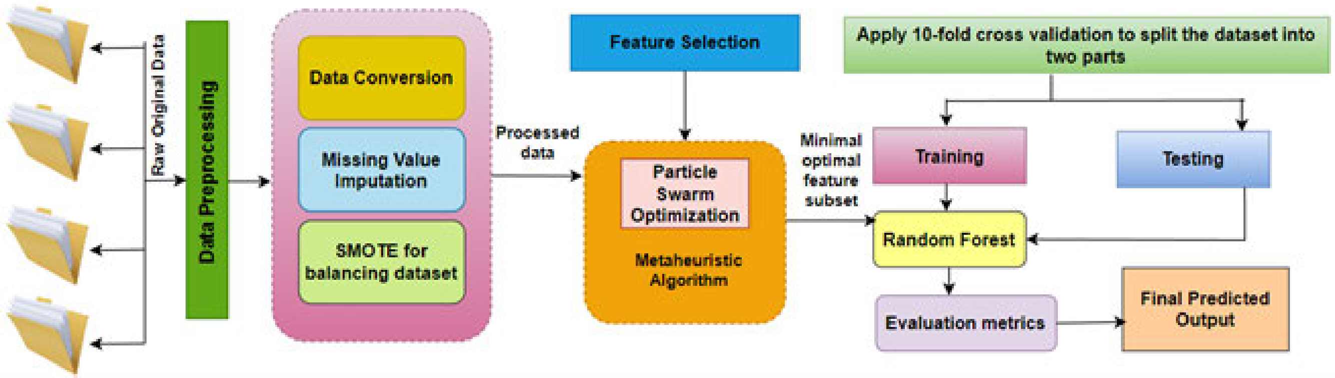


FIGURE 3. Overview of the proposed PSO-RF approach. The PSO-RF consists of a preprocessing module, a metaheuristic feature selector, and an ensemble Random Forest classifier.

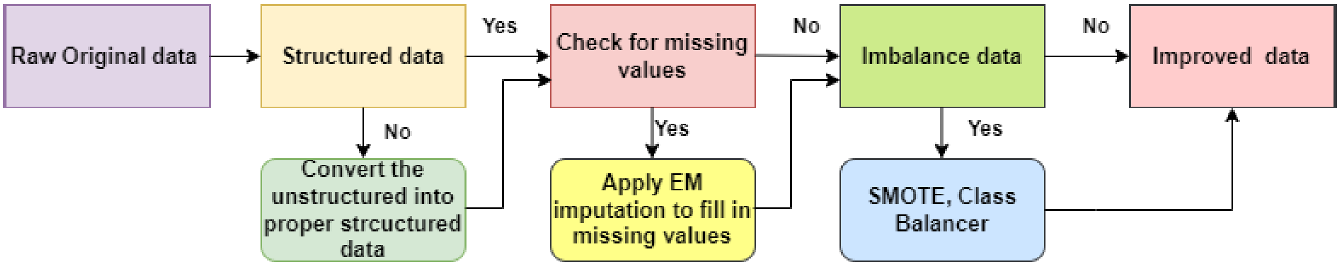


FIGURE 4. Representation of multiple stages of preprocessor module for treating raw unstructured data.

TABLE 6. Increment in the number of instances after applying SMOTE across all the datasets.

Datasets	Number of instances Before Preprocessing	Number of instances After Preprocessing
D2 (WBCD)	569	781
D3 (Z-Alizadehsani)	303	477
D4 (Exasens)	399	638
D5 (Diabetes)	390	578

Expected Minimization (EM) Imputation technique to fill in the missing values. Furthermore, all the datasets employed in this study have an imbalanced distribution of instances among different classes. To tackle this problem, SMOTE [19], [21] technique has been utilized for datasets D2, D3, D4, and D5. It creates synthetic examples of the minority class instances using the K-nearest neighbor. After the application of SMOTE filter, the rise in the number of instances can be seen in Table 6.

However, in the case of dataset D1, the distribution of instances is highly skewed The majority class (COPD) has 8723 instances and the minority class (Asthma) has 11 instances. Similarly, for other classes, the number of instances is much less as compared to the majority class. Increasing the number of instances through over-sampling using SMOTE will escalate the total number of instances to approximately 70k, quite high to be handled by the model. With such a large number of instances, the probability of getting highly noisy data is also high. Therefore, the authors utilized the Class Balancer filter to equally assign weights to all the classes as shown in Table 7.

TABLE 7. Balancing the weights of an imbalanced ICBHI dataset D1 using classbalancer.

Class Labels	Weights of Class instances before Preprocessing	Weights of Class instances after preprocessing
URTI	253	1265
LRTI	22	1265
COPD	8723	1265
Asthma	11	1265
Bronchiectasis	176	1265
Bronchiolitis	143	1265
Pneumonia	407	1265
Healthy	385	1265

The Class Balancer filter has reassigned equal weights to different class instances in such a way that the total sum of the instance weights i.e., 10120 remains the same even after balancing them. This allows the Classifier to know that each class holds equal importance and need not to be ignored.

2) STAGE 2: PARTICLE SWARM OPTIMIZATION

The second phase is the Feature selection (FS) phase which deals with selecting the best features subset that can aid in achieving optimal results. This is an optional step as it is not always required. However, FS is crucial when dealing with

Algorithm 1 Particle Swarm Optimization Based Random Forest Approach

```
Require:A Training set $S = (p_1, q_1)\dots.(p_n, q_m)$, Feature set $F$, number of trees in forest= $B$, Generation counter ($t = 1$), $T$:Maximum generators
Ensure:An optimal feature set ($F^i$), Output= $H$: predicted disease
PSO(F)
{Initialization of PSO parameters}
foreach particle $i \in 1 \dots\dots N_m$ do
    Position = $X_i(0)$
    Velocity = $V_i(0)$
    $pbest = X_i(0)$
    $gbest \leftarrow$ best of pbest
end
{Update pbest and gbest of each particle}
while$t < T$
    if$f(X_i) < f(pbest_i)$
    then
        $pbest_i(t) = X_i(t)$
        $gbest_i(t) \in \{pbest_1(t),\dots.pbest_m(t)\}|f(gbest_i(t)) = min\{f(pbest_1(t),\dots.pbest_m(t))$
    end
    for$i = 1; i \le N; i++$ do
        {Update Velocity and Position}
        $V_i(t+1) = wV_i(t) + c_1r_1(pbest_i(t) - X_i(t)) + c_2r_2(gbest_i(t) - X_i(t))$
        $X_i(t+1) = X_i(t) + V_i(t+1)$
        Evaluate fitness function of $X_i(t+1)$
    $t = t + 1$
    return $F^i$
end
RandomForest(SF^i)
$O \leftarrow \phi$
for$i = 1; i \le N; i++$ do
    $S^i \leftarrow$ A random sample from $S$
    $o_i \leftarrow RandTree(SF^i)$
    $O \leftarrow O \cup \{o_i\}$
end
return $O$
RandTree(SF)
foreach node
    $s_f \leftarrow$ a small subset of $F^i$
    Split on best features of $F^i$
return $H$
```

Chronic disease metadata as the diagnosis of a disease is done using the differential diagnosis method where the idea is to rule out the non-related diseases. Hence, a lot of tests such as laboratory tests, scans, X-rays, and blood tests were done, all of which are not really required, and also may not be related to the actual disease. And this unrelated existence of these tests might cause an overfitting problem [34]. Therefore, FS is essential before training the classification model as it will lead to a faster, more accurate, and cost-effective model.

For this purpose, this study has utilized a metaheuristic approach PSO introduced by Kennedy and Eberhart [48]. It is a stochastic population-based approach influenced by fish schooling or bird flocking behavior. It is different from other optimization algorithms like Differential Evolution in terms that it does not depend on any gradient or differential gradient. It simply explores and exploits the search space using the particle's position and velocity information. There are various advantages of PSO including being computationally inexpensive, having low system requirements, faster convergence, easy implementation, etc [49]. It is mostly used for finding the maxima or minima of a function defined over a multidimensional vector space. It performs feature selection by considering the features as particles in a high dimensional space where each particle in the swarm is an optimal solution. The fitness function is calculated for each particle in the swarm based on its position [13], [16], [19]. Each particle's position is represented as $X_i = x_{i1}, x_{i2}, \ldots\ldots\ldots x_{id}$, where d denotes the dimension. Likewise, every particle has an associated velocity, denoted by $V_i = v_{i1}, v_{i2}, \ldots\ldots\ldots, v_{id}$. After each iteration, the velocity and position values at any time instant t and $t+1$ for each particle are updated as shown in (2) and (3) respectively.

$$\begin{aligned} V_i(t+1) &= wV_i(t) + c_1 r_1(pbest_i(t) - X_i(t)) \\ &\quad + c_2 r_2(gbest_i(t) - X_i(t)) \end{aligned} \tag{2}$$

$$X_i(t+1) = X_i(t) + V_i(t+1) \tag{3}$$

In the above equations, w is the inertia constant with values between 0 and 1. It determines how much each particle keeps up with its previous velocity. In the same way, r_1 and r_2 are constants selected at random, with a value ranging from 0 to 1. Meanwhile, c_1 and c_2 are coefficients linked to cognitive and social aspects. They control the trade-off between exploration and exploitation as c_1 helps in finding the local minima and c_2 helps in finding the global minima among the local minima. The determination of the optimal local and global value is based on the variables *pbest* and *gbest* respectively. These variables depend on the position of the particle $X_i(t)$ as shown in (4) and (5). In order to determine the *pbest* and *gbest* values, the fitness function (f) of a particle at $t+1$ instant is compared with its fitness function at t instant of time.

$$pbest_i(t) = X_i(t) iff(X_i) < f(pbest_i) \tag{4}$$

$$\begin{aligned} Also, gbest_i(t) &\in \{pbest_1(t), \ldots .pbest_m(t)\} \\ |f(gbest_i(t)) &= min\{f(pbest_1(t), \ldots .pbest_m(t))\} \end{aligned} \tag{5}$$

The complete procedure for the proposed approach has been illustrated in Algorithm 1, where the preprocessed training set $S = (p_1, q_1), \ldots\ldots\ldots(p_n, q_m)$ consisting of n rows and m columns considered in this study where $S \in D$, i.e., S could be any of the five datasets D. The selected optimal feature set F^i was then passed to the training model Random forest. The goal was to select the feature set that maximizes the classification accuracy and minimizes the number of features. To achieve this goal, the fitness function (f) set for PSO is shown in (6).

$$Fitness(f) = \theta * acc(f) + (1-\theta) * (1 - \frac{N_s}{N_f}) \tag{6}$$

where N_s and N_f define the number of selected and total number of features respectively. The classification accuracy has been denoted by $acc(f)$, and θ signifies the weighing factor between the classification accuracy and the number of selected features.

3) STAGE 3: TRAINING ON RANDOM FOREST CLASSIFIER

This study has utilized a Random Forest classifier, an ensemble technique for the classification of CDs as shown in Algorithm 1. The basic idea of RF is to form a single strong classifier by combining multiple decision trees by either taking the average of their outputs or taking the majority vote. In previous works, RF has shown an excellent performance as compared to other classifiers [12], [15]. The reason is that it uses bagging for the ensemble process which reduces the correlation between the trees. Also, the variance and overfitting of the classifier get reduced [20], [31]. Moreover, by restricting the features, the decision trees can learn faster and hence can be built in a small amount of time.

TABLE 8. Description of different evaluation metrics utilized in this study.

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
Kappa statistic	$\frac{P_r(a)-p_r(e)}{1-p_r(e)}$
ROC	$\frac{TPR}{FPR}$
F-measure	$\frac{2*Recall*Precision}{Recall+Precision}$
Matthew's correlation coefficient (MCC)	$\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n} \lvert E_i - O_i \rvert$
Root Mean Square Error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i - O_i)^2}$

The algorithm 1 also considers a forest L comprising of various small decision trees l wherein for each l belonging to L, it selects a bootstrap sample S* from S. Furthermore, for each node of the tree, a very small feature set s_f is obtained from F which is then used for node splitting.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental work conducted on the four chronic disease datasets, namely D1, D2, D3, D4, and D5 has been thoroughly discussed in this section. The experiments illustrate the efficacy of the components of the proposed model by comparing them with the conventional feature selection and classification methods. Moreover, Friedman's test has also been employed as a statistical test for validating the performance of the proposed approach against previous methods.

A. EXPERIMENTAL SETUP

All experiments were run on a Windows 11 with AMD Ryzen 5 4600H with Radeon Graphics processor and 24 GB RAM. All the computations in this study have been done using three different software. The preprocessing and classification have been done using the Weka and Jupyter Notebook. In addition, for statistical testing, the SPSS tool has been utilized.

B. EVALUATION METRICS

The various evaluation metrics utilized in this study for the classification of CDs have been described in Table 8.

TABLE 9. Value of parameters set for Genetic, PSO, Firefly, and Bat algorithm across all datasets.

Genetic Parameters	Values	PSO Parameters	Values	Firefly Parameters	Values	Bat Parameters	Values
Population Size (in samples)	20	Population Size (in samples)	20	Population Size (in samples)	20	Population Size (in samples)	20
Number of iterations	20	Number of iterations	20	Number of iterations	20	Number of iterations	20
Seed	1	Seed	1	Seed	1	Seed	1
Probability of Crossover	0.6	Mutataion type	bit-flip	Mutataion type	bit-flip	Mutataion type	bit-flip
Mutation probability	0.033	Mutation Probability	0.01	Mutation Probability	0.01	Mutation Probability	0.01
		Inertia Weight	0.33	Chaotic coefficient	0.4	Chaotic coefficient	0.4
		Social weight	0.33	Absorption coefficient	0.001	frequency	0.5
		Individual weight	0.34	betaMin	0.33	loudness	0.5

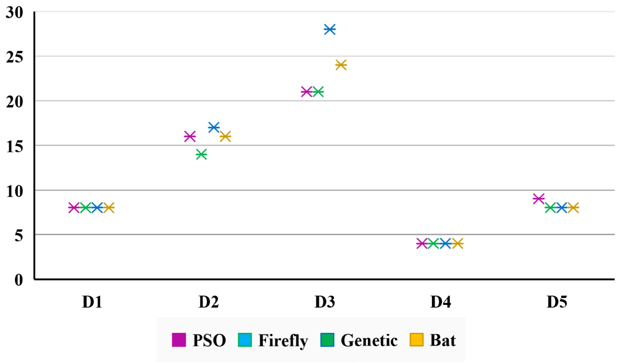


FIGURE 5. Representation of a minimal optimal number of features selected by PSO, GA, Bat,and FA techniques corresponding to datasets D1, D2, D3, D4, and D5.

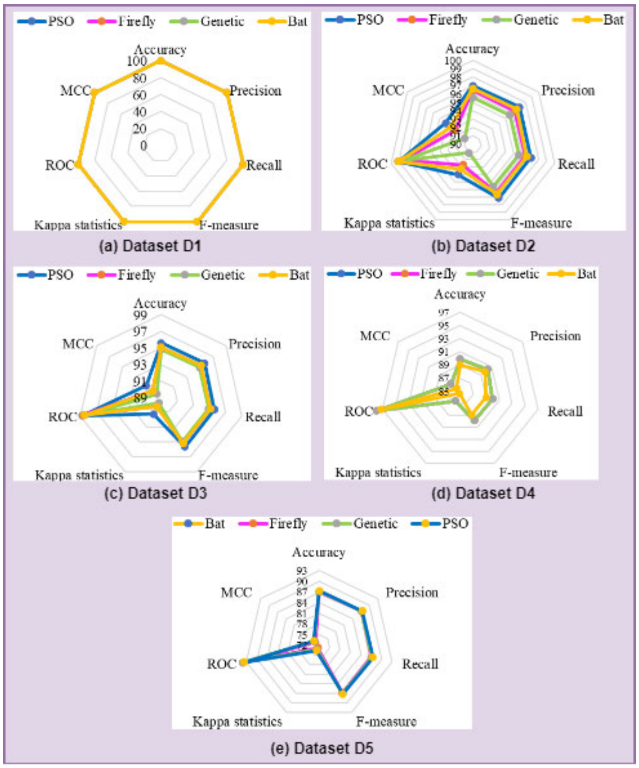


FIGURE 6. Classification output parameters of Naïve Bayes corresponding to different FS algorithms for datasets a) D1 (ICBHI), b) D2 (WBCD), c) D3 (Z-Alizadehsani), d) D4 (Exasens), and D5 (Diabetes).

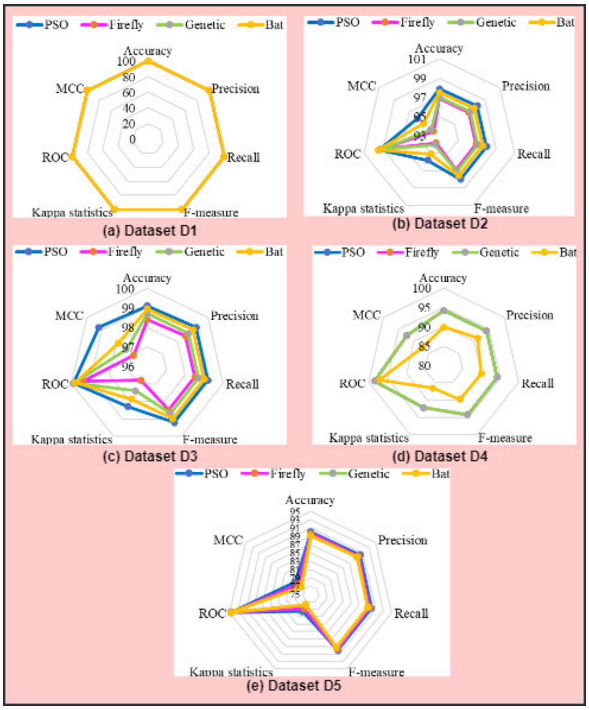
In the above Table, TP, TN, FP, and FN denote True Positive, True Negative, False Positive, and False Negative respectively. Similarly, TPR and FPR indicate a True positive rate and a False positive rate respectively. Furthermore, for MAE, n is the total number of samples, E_i is the expected or actual value, and O_i is the observed value i.e., the predicted value of i_{th} data sample obtained by the classifier. For Kappa statistics, $P_r(a)$ and $P_r(e)$ denote the actual and observed accuracy respectively.

C. COMPARISON OF PSO WITH OTHER OPTIMIZATION TECHNIQUES

This section discusses the effectiveness of the PSO optimization technique by comparing its performance with other state-of-the-art optimization feature selection methods Genetic Algorithm (GA), Bat and Firefly Algorithm (FA). In this regard, the parameters corresponding to PSO, GA, Bat, and FA have been set across all five datasets for determining the minimal optimal feature subset as shown in Table 9.

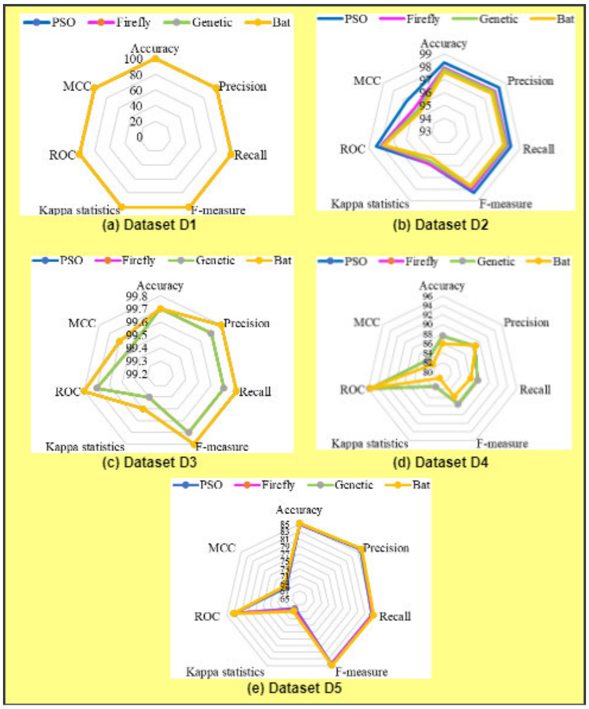
The number of minimal attributes resulting from all four optimization techniques are shown in Figure 5.

Different techniques provided the minimal set of features across all the datasets except D1 as it already contained the minimal attributes. As a rule of thumb, it is known that neither too many nor too few features should be utilized for the prediction [25]. This study utilized the original set of features for D1 as the resulting optimal features were too less. For dataset D3, PSO and FA has provided the minimal set of features and for dataset D2, FA provided the minimal set. However, for dataset D4 all three techniques resulted in a minimal subset of features. Also, for dataset D5, PSO obtained a minimal optimal feature set of 9 attributes which is higher than the feature set provided by the other FS techniques. Furthermore, to validate the performance of PSO over GA, FA, and Bat algorithms, Radar charts have been



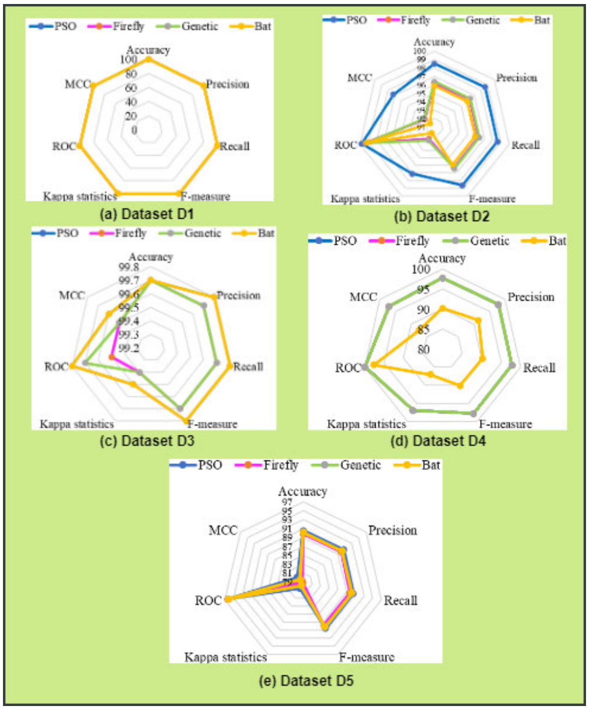
(a) Dataset D1
(b) Dataset D2
(c) Dataset D3
(d) Dataset D4
(e) Dataset D5

FIGURE 7. Classification output parameters of MLP corresponding to different FS algorithms for datasets a) D1 (ICBHI), b) D2 (WBCD), c) D3 (Z-Alizadehsani), d) D4 (Exasens), and D5 (Diabetes).



(a) Dataset D1
(b) Dataset D2
(c) Dataset D3
(d) Dataset D4
(e) Dataset D5

FIGURE 8. Classification output parameters of SMO corresponding to different FS algorithms for datasets a) D1 (ICBHI), b) D2 (WBCD), c) D3 (Z-Alizadehsani), d) D4 (Exasens), and D5 (Diabetes).



(a) Dataset D1
(b) Dataset D2
(c) Dataset D3
(d) Dataset D4
(e) Dataset D5

FIGURE 9. Classification output parameters of Bagging corresponding to different FS algorithms for datasets a)D1 (ICBHI), b) D2 (WBCD), c) D3 (Z-Alizadehsani), d) D4 (Exasens), and D5 (Diabetes).

drawn across all five datasets for different ML classifiers. Each chart evaluates the performance of all optimization techniques for different evaluation metrics. Firstly, the FS techniques have been compared for the Naïve Bayes classifier as shown in Figure 6.

It can be clearly observed from the above figure that for datasets D2, D3, and D5, PSO has shown the best performance. However, in the case of datasets D1 and D4, a similar number of attributes has been obtained by all the FS techniques, consequently leading to overlapping charts. Similarly, for MLP, SMO, Bagging, and RF, different charts have been obtained as shown in Figure 7, 8, 9, and 10 respectively.

It is clear from the figures above that the minimal optimal feature set obtained from PSO has greatly helped all the classifiers in achieving the highest performance as compared to other FS techniques.

D. COMPARISON OF RF CLASSIFIER WITH BENCHMARK ML CLASSIFIERS

This section benchmarks the performance of the ensemble RF classifier towards other state-of-the-art classifiers, i.e., NB, MLP, SMO, and Bagging. In this regard, the hyperparameters corresponding to all these classifiers have been set across all five datasets as shown in Table 10.

Moreover, this study has employed 10-fold cross-validation for splitting the dataset into training and testing sets. Thereafter, the training set and selected feature set were passed to all the classifiers for the classification of CDs. The resulting classification performance of all the classifiers has been compared across all the datasets for different evaluation metrics as shown in Table 11 and 12.

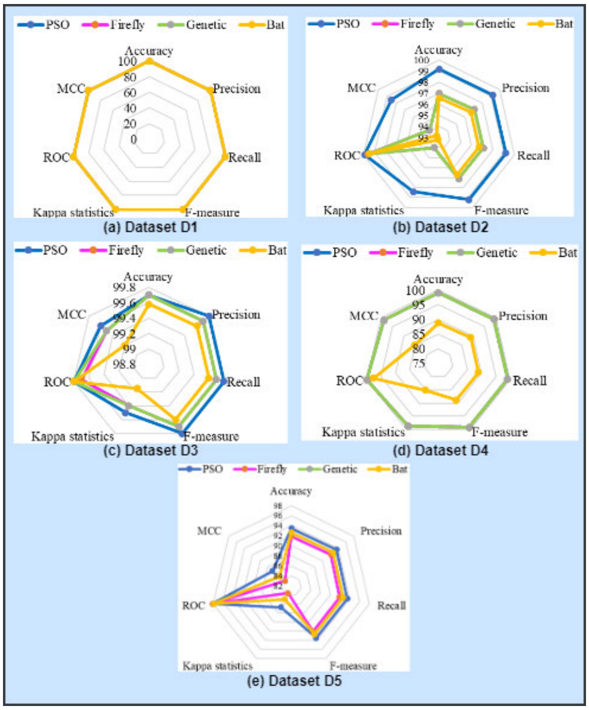


FIGURE 10. Classification output parameters of RF corresponding to different FS algorithms for datasets a) D1 (ICBHI), b) D2 (WBCD), c) D3 (Z-Alizadehsani), d) D4 (Exasens), and D5 (Diabetes).

TABLE 10. Values of hyperparameters across all the classifiers.

Classifier	Hyperparameters
Naïve Bayes	BatchSize=100, Number of decimal places= 2, Kernel estimator= False, Supervised Discretization= False
MLP	BatchSize=100, decay= False, Hidden layers= a, learning rate= 0.3, momentum= 0.2, seed=0, training time=500, Validation threshold= 20
SMO	BatchSize=100, c=1.0, Calibrator= Logistics, epsilon= 1.0E-12, FilterType=Normalize Training data, Kernel= PolyKernel, Seed=1, Tolerance parameter= 0.001
Bagging	BatchSize=100, Classifier= REPTree, Number of iterations= 10, seed =1
Random Forest	BatchSize=100, Number of execution slot= 1, Number of iteration= 100, seed =1

The results obtained from the experimentation work illustrated two important observations.

- Firstly, a situation of accuracy paradox has been raised for dataset D1. The performance of all the classifiers for different metrics across dataset D1 is ideal, which is quite impossible. This is due to the presence of a high imbalance across the classes of dataset D1. These biased outcomes have resulted because of the biased data.
- Secondly, there are cases where multiple classifiers have shown similar results corresponding to the same metric. For example, for dataset D3, SMO, Bagging, and RF have shown similar performance in terms of Accuracy.

Hence, to further assess the classification performance of the proposed approach against other state-of-the-art classifiers, some statistical tests are required that are discussed in the subsection below.

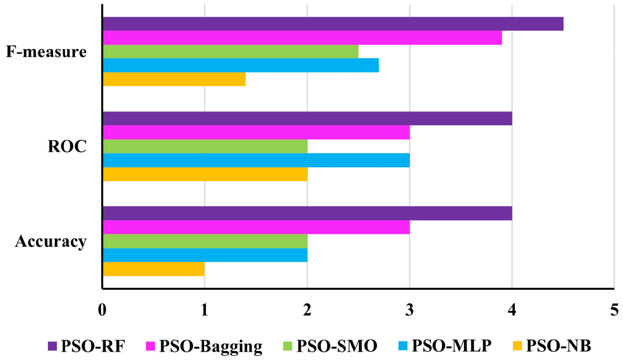


FIGURE 11. Comparison of Accuracy, ROC, and F-measure across all classifiers in terms of mean Rank calculated by Friedman's Test.

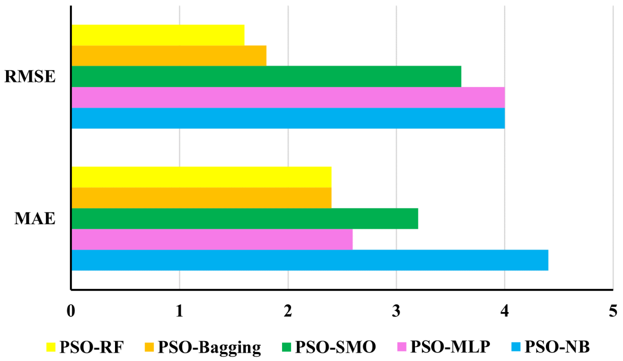


FIGURE 12. Comparison of MAE and RMSE across all classifiers in terms of mean Rank calculated by Friedman's Test.

E. STATISTICAL TESTING

In this section, a thorough comparison has been conducted between the proposed approach and other benchmark classifiers, utilizing Friedman's statistical test to determine the results [19]. This test with the associated p-value has been performed for multiple comparisons. It has been undertaken to detect the performance difference between the PSO-RF and different classifiers. The null hypothesis with threshold value $p = 0.05$ considered for this study was that there is no significant difference between PSO-RF and other classifiers. The indication of a significant difference is appraised by p<0.05. Different test statistics set for Friedmann's test have been shown in Table 13.

It is worth mentioning that the performance difference between PSO-RF and other classifiers is highly significant ($p < 0.05$) for Accuracy, F-measure, and RMSE. Hence, rejecting the null hypothesis for these parameters that, there is no significant difference between PSO-RF and other classifiers.

The Friedmann mean rank obtained on the above experimental results for different classifiers across different evaluation metrics is shown in Figures 11 and 12. In terms of Accuracy, ROC, and F-measure, the higher the rank of the classifier the better the classifier. Whereas for MAE, and RMSE, the lower the error rank the better the classifier.

TABLE 11. Comparison of performance of classifiers across all five datasets (D1, D2, D3, D4, and D5) in terms of accuracy (in %), ROC (in %), F-measure (in %).

Evaluation metrics	Datasets	PSO-NB	PSO-MLP	PSO-SMO	PSO-Bagging	PSO-RF
Accuracy	D1	99.6	100	100	100	100
	D2	95.6	97.9	98.3	98.5	99.2
	D3	97	99.1	99.7	99.7	99.7
	D4	89.9	94.2	87.6	97.8	99.0
	D5	87.3	90.0	84.9	90.5	93.5
ROC	D1	100	100	100	100	100
	D2	98.8	99.4	98.4	99.8	99.9
	D3	99.1	99.9	99.8	99.7	99.7
	D4	98	98.4	95.7	99.9	99.9
	D5	93.6	94.7	82.9	96.3	98.1
F-measure	D1	99.6	100	100	100	100
	D2	95.6	98	98.3	98.6	99.2
	D3	97.1	99.2	99.8	99.8	99.8
	D4	89.9	94.2	87.5	97.8	99.1
	D5	87.2	90.0	84.5	90.5	93.5

TABLE 12. Comparison of performance of classifiers across all five datasets (D1, D2, D3, D4, and D5) in terms of MAE and RMSE.

Evaluation metrics	Datasets	PSO-NB	PSO-MLP	PSO-SMO	PSO-Bagging	PSO-RF
MAE	D1	0.027	0.018	0.1875	0.0011	0.0083
	D2	0.0421	0.0235	0.0166	0.0392	0.0192
	D3	0.0385	0.0121	0.0021	0.006	0.054
	D4	0.072	0.0437	0.2603	0.0218	0.0135
	D5	0.1516	0.1046	0.1509	0.1402	0.1231
RMSE	D1	0.0768	0.312	0.2912	0.0011	0.0083
	D2	0.1877	0.1362	0.129	0.1132	0.075
	D3	0.1622	0.907	0.0458	0.0476	0.1004
	D4	0.1934	0.1567	0.3279	0.0921	0.0727
	D5	0.3173	0.2905	0.3884	0.2637	0.2284

TABLE 13. Values of different test statistics are set across different performance metrics during friedman's test.

Statistics	Accuracy	ROC	F-measure	MAE	RMSE
Number of datasets	5	5	5	5	5
Chi-square value	13.860	9.026	13.860	5.760	11.520
Degree of freedom	4	4	4	4	4
Significance	0.008	0.060	0.008	0.218	0.021

The results obtained from Friedman's Test showed that out of all the classifiers, PSO-RF obtained the highest mean rank in terms of accuracy, ROC, and F-measure. Also, the mean of MAE is similar for PSO-Bagging and PSO-RF. Similarly, the mean RMSE is lowest for PSO-RF. Hence, it is evidently visible from Fig. 11 and 12 that the proposed model (PSO-RF) has risen as the best model as it exhibits the highest ranks among all the classifiers.

F. COMPARISON OF PROPOSED APPROACH WITH PREVIOUS METHODS

For the sake of universality and comprehensiveness, this study further contrasts the proposed approach with other existing studies. A state-of-art comparison with the proposed approach for dataset D1 has also been shown in Figure 13.

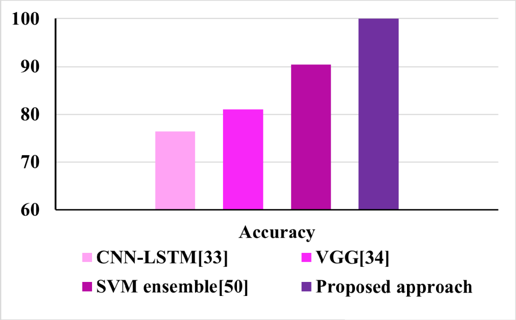


FIGURE 13. Comparison of Proposed approach with previous studies with respect to ICBHI Dataset in terms of Accuracy. Convolutional Neural Network-Long short term memory (CNN-LSTM), Visual Geometry Group (VGG).

From Fig. 13, it is evident that for Datasets D1, the proposed approach, i.e., PSORF has outperformed the previous studies' results by obtaining the highest accuracy of 100%. The accuracies obtained by studies [33], [34], [50] were way too low for dataset D1 as compared to the proposed approach. The second highest accuracy was obtained in [50] wherein the author utilized an ensemble of Support vector machines (SVM). However, the study couldn't identify the feature importance as it utilized the radial basis function

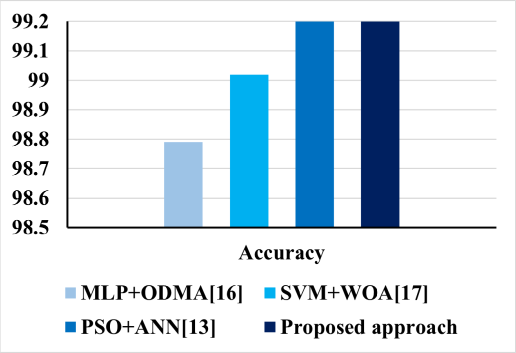


FIGURE 14. Comparison of Proposed approach with previous studies with respect to WBCD Dataset in terms of Accuracy.Multilayer perceptron+Open source development Model Algorithm (MLP+ODMA), Support Vector Machine- Wolf Optimization Algorithm (SVM+WOA),Particle Swarm Optimization+Artificial Neural Network (PSO+ANN).

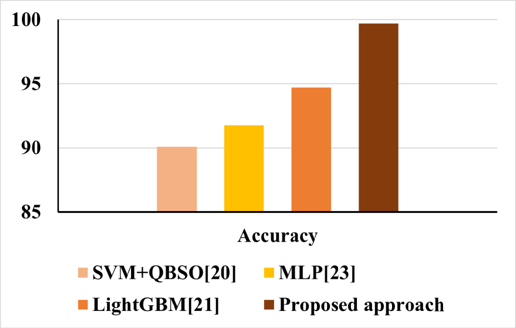


FIGURE 15. Comparison of Proposed approach with previous studies with respect to Z-Alizadehsani Dataset in terms of Accuracy.Support Vector Machine+Q learning based Bee Swwarm Optimization (SVM+QBSO).

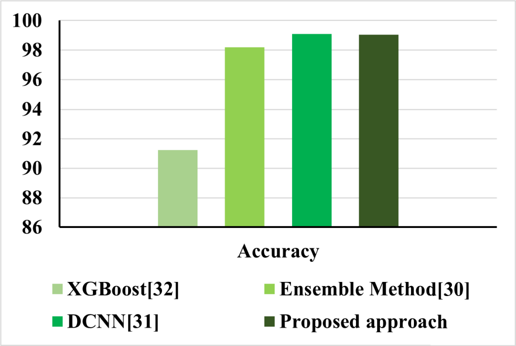


FIGURE 16. Comparison of Proposed approach with previous studies with respect to Exasens Dataset D4 in terms of Accuracy. Deep Convolution neural network (DCNN).

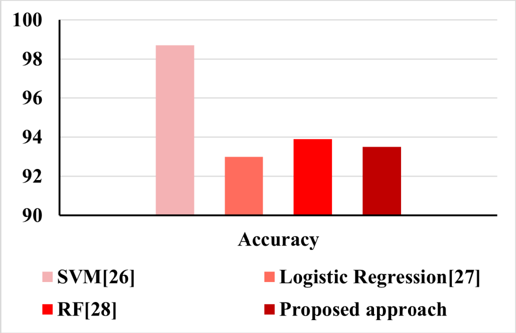


FIGURE 17. Comparison of Proposed approach with previous studies with respect to Vanderbilt Diabetes Dataset D5 in terms of Accuracy.

to derive the best-performing model. Hence, it could be said that in terms of feature importance and classification performance, the proposed approach performed the best for dataset D1. For dataset D2 as can be seen in Figure 14, the proposed approach obtained similar results as that of study [13] in terms of accuracy.

Howbeit, this might be due to overfitting as the dataset was left imbalanced in study [13], and also the researchers utilized a highly computational Deep learning model for obtaining high accuracy. Similarly, for studies [16], [17], the dataset was left imbalanced, and no statistical tests were performed to support the classification performances obtained by their respective models. Therefore, in terms of computational power, the proposed approach for dataset D2 is better than all previous studies. Furthermore, for dataset D3, it is clearly evident from the above Figure 15, that the proposed approach obtained the highest accuracy of 99.7% as compared to studies [20], [21], [23].

The second highest accuracy has been obtained by study [21] wherein the authors utilized the LightGBM model for the detection of CAD disease. However, the problem with previous studies related to dataset D3 had some limitations such as increased time complexity for study [20], no feature selection in study [21], and imbalance data problem in study [23]. All these limitations have been overcome in the proposed approach, consequently leading to better accuracy as compared to the previous studies. Similarly, in the case of dataset D4 as shown in Figure 16, the proposed approach obtained the second-highest accuracy of 99.05% which is 0.5% less than the accuracy obtained by researcher [31].

The proposed approach has completely rectified the problem of missing value by utilizing EM Imputation whereas the problem still persists in study [31], [32]. At last, for dataset D5, the proposed approach obtained the third highest accuracy of 93.5% as shown in Figure 17.

The other studies [26], [27] obtained an accuracy of 98.7% and 93.9% respectively. However, the problem with these approaches is that the dataset had missing values and was left imbalanced. In addition, the study [26] utilized GWO and WOA (that should not be used as these algorithms exhibit center bias problem) as the base feature selection technique as their proposed model. The technique proposed in this study is free from center bias problem and also the problem of missingness and imbalance data has been rectified. Hence, at last, it could be said that across all the datasets except D5, the proposed approach, i.e., PSORF has performed the best. It has not only detected Chronic diseases but also multi-classified symptomatically similar diseases. This study also has some limitations. Firstly, for dataset D1, the proposed approach obtained almost ideal results which might be a result of the presence of a high imbalance in the dataset. The ML classifiers utilized in this study are not complex enough to deal with such highly imbalanced data. Secondly. to tackle down the imbalance data problem, this study has utilized SMOTE filter which might result in the generation of some noisy data. Therefore, in the future, this study aims to provide a suitably complex AI-based predictive model for the multi-classification of diseases in dataset D1. Furthermore, for the problem of imbalance dataset, different variants of SMOTE can be applied in the future studies.

V. CONCLUSION

This study aimed to provide an efficient Machine learning framework PSORF that can not only detect but also Classify similar Chronic diseases such as COPD, Asthma, Bronchiectasis, etc. For this purpose, this study considered five different datasets across which a series of experiments have been performed. The datasets obtained from public repositories suffered from missing values and imbalanced data problems that were rectified through EM Imputation and SMOTE techniques. The processed data was then passed through the PSO-RF framework which provided the best optimal feature set and efficient classification result on all the datasets. In addition, to validate the classification performance of the PSORF framework, both PSO and RF were compared with different metaheuristic and ML classifiers respectively. The performance of PSO with other metaheuristic techniques, namely firefly, Bat and Genetic search were compared through radar graphs on the basis of various evaluation metrics. It was evident from the graphs that across all the datasets, PSO provided the best results. Hence, for further evaluation, five different PSO-based classifiers were compared by using various performance metrics. The results showed that among all the classifiers, the PSO-based RF classifier outperformed the other classifiers in terms of Accuracy, F-measure, and ROC. However, there were some classifiers whose performances were similar across all the datasets. Therefore, for further clarification on the classification performance of the classifiers, Friedman's testing was performed. The test results proved that among all the classifiers PSO-RF achieved the highest rank indicating that it has outperformed other classifiers. The proposed PSO-RF framework not only classified the binary Chronic diseases such as Breast cancer, Diabetes and Heart disease but also classified multiple chronic diseases that were symptomatically similar such as COPD, Asthma, Pneumonia, Bronchiectasis, etc.

REFERENCES

[1] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, A. Alonso, A. Z. Beaton, M. S. Bittencourt, A. K. Boehme, A. E. Buxton, A. P. Carson, Y. Commodore-Mensah, and M. S. Elkind, "Heart disease and stroke statistics, 2022 update: A report from the American heart associatio," *Circulation*, vol. 145, no. 8, pp. e153–e639, 2022.

[2] A. Singh, N. Prakash, and A. Jain, "A review on prevalence of worldwide COPD situation," in *Proceedings of Data Analytics and Management* (Lecture Notes in Networks and Systems), vol. 572, A. Khanna, Z. Polkowski, and O. Castillo, Eds. Singapore: Springer, 2023.

[3] L. J. Grimm, C. S. Avery, E. Hendrick, and J. A. Baker, "Benefits and risks of mammography screening in women ages 40 to 49 years," *J. Primary Care Community Health*, vol. 13, Jan. 2022, Art. no. 215013272110583.

[4] S. Selvakani, K. Vasumathi, and V. Aadhiseshan, "Application of machine learning in predicting heart disease," *Asian Basic Appl. Res. J.*, vol. 5, pp. 61–68, Apr. 2023.

[5] A. Chaurasia, "Ensemble technique to predict heart disease using machine learning classifiers," *Netw. Biol.*, vol. 13, no. 1, p. 1, 2023.

[6] G. N. Ahamad, Shafiullah, H. Fatima, Imdadullah, S. M. Zakariya, M. Abbas, M. S. Alqahtani, and M. Usman, "Influence of optimal hyperparameters on the performance of machine learning algorithms for predicting heart disease," *Processes*, vol. 11, no. 3, p. 734, Mar. 2023.

[7] A. Singh and N. Prakash, "A review of AI models for prediction and detecting heart disease for improved wellbeing," *Vivekananda J. Res.*, vol. 10, pp. 14–25, Oct. 2021.

[8] S. W. Ali, M. Asif, M. Rashid, S. Tanvir, S. Shams, and S. Abid, "Detection of crackle and wheeze in lung sound using machine learning technique for clinical decision support system," *Vawkum Trans. Comput. Sci.*, vol. 11, no. 1, pp. 67–78, 2023.

[9] M. A. Elsadig, A. Altigani, and H. T. Elshoush, "Breast cancer detection using machine learning approaches: A comparative study," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 1, p. 736, Feb. 2023.

[10] V. R. Allugunti, "Breast cancer detection based on thermographic images using machine learning and deep learning algorithms," *Int. J. Eng. Comput. Sci.*, vol. 4, no. 1, pp. 49–56, Jan. 2022.

[11] B. S. Abunasser, M. R. J. Al-Hiealy, I. S. Zaqout, and S. S. Abu-Naser, "Breast cancer detection and classification using deep learning Xception algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 7, pp. 223–228, 2022.

[12] T. O. Oladele, B. J. Olorunsola, T. O. Aro, H. B. Akande, and O. A. Olukiran, "Nature-inspired meta-heuristic optimization algorithms for breast cancer diagnostic model: A comparative study," *FUOYE J. Eng. Technol.*, vol. 6, no. 1, pp. 26–29, Mar. 2021.

[13] R. O. Ogundokun, S. Misra, M. Douglas, R. Damaševičius, and R. Maskeliūnas, "Medical Internet-of-Things based breast cancer diagnosis using hyperparameter-optimized neural networks," *Future Internet*, vol. 14, no. 5, p. 153, May 2022.

[14] B. Sahu, S. Mohanty, and S. Rout, "A hybrid approach for breast cancer classification and diagnosis," *ICST Trans. Scalable Inf. Syst.*, vol. 6, no. 20, Jul. 2018, Art. no. 156086.

[15] B. J. Olorunsola, T. O. Oladele, T. O. Aro, H. Babalola, and O. A. Olukiran, "Performance comparison of selected swarm intelligence algorithms on breast cancer diagnosis," *Afr. J. MIS*, vol. 3, no. 1, pp. 5–21, 2021.

[16] Z. Guo, L. Xu, and N. A. Asgharzadeholiaee, "A homogeneous ensemble classifier for breast cancer detection using parameters tuning of MLP neural network," *Appl. Artif. Intell.*, vol. 36, no. 1, Dec. 2022, Art. no. 2031820.

[17] X. Jia, X. Sun, and X. Zhang, "Breast cancer identification using machine learning," *Math. Problems Eng.*, vol. 2022, pp. 1–8, Oct. 2022.

[18] H. Huang, X. Feng, S. Zhou, J. Jiang, H. Chen, Y. Li, and C. Li, "A new fruit fly optimization algorithm enhanced support vector machine for diagnosis of breast cancer based on high-level features," *BMC Bioinf.*, vol. 20, no. S8, pp. 1–14, Jun. 2019.

[19] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "C-CADZ: Computational intelligence system for coronary artery disease detection using Z-Alizadeh Sani dataset," *Int. J. Speech Technol.*, vol. 52, no. 3, pp. 2436–2464, Feb. 2022.

[20] Y. A. Z. A. Fajri, W. Wiharto, and E. Suryani, "Hybrid model feature selection with the bee swarm optimization method and Q-learning on the diagnosis of coronary heart disease," *Information*, vol. 14, no. 1, p. 15, Dec. 2022.

[21] S. Zhang, Y. Yuan, Z. Yao, J. Yang, X. Wang, and J. Tian, "Coronary artery disease detection model based on class balancing methods and LightGBM algorithm," *Electronics*, vol. 11, no. 9, p. 1495, May 2022.

[22] J. Hassannataj Joloudari, F. Azizi, M. A. Nematollahi, R. Alizadehsani, E. Hassannatajjeloudari, I. Nodehi, and A. Mosavi, "GSVMA: A genetic support vector machine ANOVA method for CAD diagnosis," *Frontiers Cardiovascular Med.*, vol. 8, p. 2178, Feb. 2022.

[23] B. Kolukisa and B. Bakir-Gungor, "Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis," *Comput. Standards Interfaces*, vol. 84, Mar. 2023, Art. no. 103706.

[24] B. Kolukisa, L. Yavuz, A. Soran, B.-G. Burcu, D. Tuncer, A. Onen, and V. C. Gungor, "Coronary artery disease diagnosis using optimized adaptive ensemble machine learning algorithm," *Int. J. Bioscience, Biochemistry Bioinf.*, vol. 10, no. 1, pp. 58–65, 2020.

[25] A. Singh and A. Payal, "CAD diagnosis by predicting stenosis in arteries using data mining process," *Intell. Decis. Technol.*, vol. 15, no. 1, pp. 59–68, Mar. 2021.

[26] S. Amutha and J. R. Sekar, "An optimized framework for diabetes mellitus diagnosis using grid search based support vector machine," in *Proc. Int. Conf. Comput., Commun., Signal Process.* Cham, Switzerland: Springer, Jan. 2023, pp. 153–167.

[27] A. C. Ramachandra and D. Murthy, "Diabetes prediction using machine learning approach," *Strad Res.*, vol. 10, no. 8, 2023.

[28] S. Gill and P. Pathwar, "Prediction of diabetes using various feature selection and machine learning paradigms," in *Modern Approaches in Machine Learning & Cognitive Science: A Walkthrough*. Cham, Switzerland: Springer, 2022, pp. 133–146.

[29] P. Rajendra and S. Latifi, "Prediction of diabetes using logistic regression and ensemble techniques," *Comput. Methods Programs Biomed. Update*, vol. 1, Jan. 2021, Art. no. 100032.

[30] J. Dhar, "Multistage ensemble learning model with weighted voting and genetic algorithm optimization strategy for detecting chronic obstructive pulmonary disease," *IEEE Access*, vol. 9, pp. 48640–48657, 2021.

[31] R. R. Irshad, S. Hussain, S. S. Sohail, A. S. Zamani, D. Ø. Madsen, A. A. Alattab, A. A. A. Ahmed, K. A. A. Norain, and O. A. S. Alsaiari, "A novel IoT-enabled healthcare monitoring framework and improved grey wolf optimization algorithm-based deep convolution neural network model for early diagnosis of lung cancer," *Sensors*, vol. 23, no. 6, p. 2932, Mar. 2023.

[32] P. S. Zarrin, N. Roeckendorf, and C. Wenger, "In-vitro classification of saliva samples of COPD patients and healthy controls using machine learning tools," *IEEE Access*, vol. 8, pp. 168053–168060, 2020.

[33] G. Petmezas, G.-A. Cheimariotis, L. Stefanopoulos, B. Rocha, R. P. Paiva, A. K. Katsaggelos, and N. Maglaveras, "Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function," *Sensors*, vol. 22, no. 3, p. 1232, Feb. 2022.

[34] S. W. Ali, M. Asif, M. Rashid, S. Tanvir, S. Shams, and S. Abid, "Detection of crackle and wheeze in lung sound using machine learning technique for clinical decision support system," *Vawkum Trans. Comput. Sci.*, vol. 11, no. 1, pp. 67–78, 2023.

[35] *ICBHI Dataset*. Accessed: Jun. 20, 2023. [Online]. Available: https://paperswithcode.com/dataset/icbhi-respiratory-sound-database

[36] *WBCD*. Accessed: Jun. 20, 2023. [Online]. Available: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

[37] *Z-Alizadehsani Dataset*. Accessed: Jun. 20, 2023. [Online]. Available: https://archive.ics.uci.edu/dataset/extention-of-z-alizadehsani-dataset

[38] *EXASENS*. Accessed: Jun. 20, 2023. [Online]. Available: https://archive.ics.uci.edu/dataset/523/exasens

[39] *Diabetes Prediction Dataset*. Accessed: Sep. 20, 2023. [Online]. Available: https://data.world/informatics-edu/diabetes-prediction

[40] M. Zhang, M. Li, L. Guo, and J. Liu, "A low-cost AI-empowered stethoscope and a lightweight model for detecting cardiac and respiratory diseases from lung and heart auscultation sounds," *Sensors*, vol. 23, no. 5, p. 2591, Feb. 2023.

[41] C. Wall, L. Zhang, Y. Yu, A. Kumar, and R. Gao, "A deep ensemble neural network with attention mechanisms for lung abnormality classification using audio inputs," *Sensors*, vol. 22, no. 15, p. 5566, Jul. 2022.

[42] A. Mohamed, E. Amer, S. N. Eldin, J. Khaled, and M. Hossam, "The impact of data processing and ensemble on breast cancer detection using deep learning," *J. Comput. Commun.*, vol. 1, no. 1, pp. 27–37, Feb. 2022.

[43] X. Wang, I. Ahmad, D. Javeed, S. Zaidi, F. Alotaibi, M. Ghoneim, Y. Daradkeh, J. Asghar, and E. Eldin, "Intelligent hybrid deep learning model for breast cancer detection," *Electronics*, vol. 11, no. 17, p. 2767, Sep. 2022.

[44] H. Mohammedqasim, R. Mohammedqasem, O. Ata, and E. I. Alyasin, "Diagnosing coronary artery disease on the basis of hard ensemble voting optimization," *Medicina*, vol. 58, no. 12, p. 1745, Nov. 2022.

[45] *Vanderbilt Diabetes Datasets*. Accessed: Sep. 20, 2023. [Online]. Available: https://hbiostat.org/data/

[46] J. Kudela, "The evolutionary computation methods no one should use," 2023, *arXiv:2301.01984*.

[47] Y. Wan, Z. Wang, and T.-Y. Lee, "Incorporating support vector machine with sequential minimal optimization to identify anticancer peptides," *BMC Bioinf.*, vol. 22, no. 1, p. 286, May 2021.

[48] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Netw. (ICNN)*, vol. 4, Aug. 2002, pp. 1942–1948.

[49] A. Singh and A. Jain, "Financial fraud detection using bio-inspired key optimization and machine learning technique," *Int. J. Secur. Appl.*, vol. 13, no. 4, pp. 75–90, Dec. 2019, doi: 10.33832/ijsia.2019.13.4.08.

[50] J. S. Park, K. Kim, J. H. Kim, Y. J. Choi, K. Kim, and D. I. Suh, "A machine learning approach to the development and prospective evaluation of a pediatric lung sound classification model," *Sci. Rep.*, vol. 13, no. 1, p. 1289, Jan. 2023.

AKANSHA SINGH received the B.Tech. and M.Tech. degrees in computer science and engineering from Guru Gobind Singh Indraprastha University (GGSIPU), Delhi, India, in 2017 and 2020, respectively, where she is currently pursuing the Ph.D. degree in computer science and engineering. Her research interests include machine learning, computational metaheuristic models, deep learning, bioinformatics, and data mining. She was a recipient of two best paper awards at an international and national conference respectively. Her honors include the Short Term Research Fellowship (STRF) from GGSIPU and the STEM Fellowship.

NUPUR PRAKASH received the B.E. degree in electronics and communication and the M.E. degree in computer science and technology from the University of Roorkee (now IIT Roorkee), in 1981 and 1986, respectively, and the Ph.D. degree from Punjab University, in 1998. She is currently a Professor with the Department of Computer Science and Engineering and holds the position of the Vice Chancellor of The Northcap University, Gurgaon, India. Prior to joining The NorthCap University, she was the Vice-Chancellor of Indira Gandhi Delhi Technical University for Women; the Principal of the Indira Gandhi Institute of Technology, Delhi; the Dean of the School of Engineering and Technology; and the Dean of the School of ICT, Guru Gobind Singh Indraprastha University, Government of Delhi. She has been a strong propagator of STEM education among girls and has won many awards and accolades. She has guided 12 Ph.D. scholars and authored more than 100 research papers and articles in various national and international journals/conferences of repute. Her H-index and i10 index are 17 and 30, respectively, with 1844 citations. Her research interests include artificial neural networks, natural language processing, mobile communication, secure wireless networks, and machine learning algorithms. She is a Life Member of the Computer Society of India (CSI) and a Former Member of the IEEE Women in Engineering (WIE), USA. She has chaired various expert committees of UGC, NBA, and NAAC.

ANURAG JAIN received the M.Tech. degree from IIT Kharagpur and the Ph.D. degree from Guru Gobind Singh Indraprastha University, Delhi, India.

He is currently a Professor with Guru Gobind Singh Indraprastha University. He is doing research in the areas of healthcare, cybersecurity, and speech processing. He has also been involved in identifying the importance of ML and data science in his research domain. He has published many national and international research papers in many reputed journals and conferences. His i10 index is 14 with nearly 675 citations. His research interests include speech processing, natural language processing, artificial intelligence, machine learning, and data mining in the healthcare domain.

Prof. Jain is a Life Member of the Computer Society of India (CSI).