IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 11, 2021, accepted October 19, 2021, date of publication October 26, 2021, date of current version November 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3122789

HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm

SARRIA E. A. ASHRI, M. M. EL-GAYAR, AND EMAN M. EL-DAYDAMONY
Department of Information Technology, Faculty of Computer and Information Science, Mansoura University, Mansoura 35516, Egypt
Corresponding author: M. M. EL-Gayar (mostafa_elgayar@mans.edu.eg)

ABSTRACT Supervised machine learning algorithms are powerful classification techniques commonly used to build prediction models that help diagnose the disease early. However, some challenges like overfitting and underfitting need to be overcome while building the model. This paper introduces hybrid classifiers using the ensembled model with a majority voting technique to improve prediction accuracy. Furthermore, a proposed preprocessing technique and features selection based on a genetic algorithm is suggested to enhance prediction performance and overall time consumption. In addition, the 10-folds cross-validation technique is used to overcome the overfitting problem. Experiments were performed on a dataset for cardiovascular patients from the UCI Machine Learning Repository. Through a comparative analytical approach, the study results indicated that the proposed ensemble classifier model achieved a classification accuracy of 98.18% higher than the rest of the relevant developments in the study.

INDEX TERMS Cardiovascular disease, supervised machine learning algorithms, simple genetic algorithm, ensembled model, majority voting technique.

I. INTRODUCTION

There are several cardiovascular diseases, such as heart failure, angina, cardiomyopathy, and arrhythmia. Heart disease is a universal disease that affects many people, especially during middle or old age [1]. Heart diseases are more common among men than among women. According to WHO statistics [2], it is estimated that 30% of deaths in developing countries are caused by heart disease [3], [4]. One-third of global deaths worldwide are due to heart disease [5]. Half of the deaths are in the United States, and other developed countries are due to heart disease. Every year approximately 17 million people die from cardiovascular disease (CVD) worldwide [2].

Currently, we have a wealth of big data provided by patients' electronic health records. Technology has also provided us with many methods, techniques, and models that enable data scientists and researchers to contribute to medical development. Through analytics, the data can determine the causes of the disease and the medical team's contribution by spreading community awareness through prevention. By adopting preventative behavior, a person can better avoid disease. "Prevention is better than cure". Therefore, there are many challenges associated with this field, which can be summarized as follows:

- Patient health records contain a wide variability and a diversity of features [4].
- The diagnosis of any disease depends on linking symptoms together and this depends on the speed of diagnosis in real-time. Therefore, any diagnostic system requires high speed and accuracy in performing the tasks [21].
- The classification process using machine learning algorithms may suffer from overfitting problems [10].

The classification technique is a commonly used Machine Learning (ML) application with medical diagnostics and predictions. Most of the time, classification accuracy is used to measure model performance. However, it is not enough to judge the model's accuracy. Therefore, several classification metrics have been proposed to evaluate the machine learning model to obtain a concrete evaluation of our model [6]–[8]. In this manuscript, a hybrid of five ML models was created to classify and predict CVD occurrence. These models are Logistic Regression (LR), Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Decision Tree (DT),

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

and Random Forest (RF). We adopted a set of metrics to evaluate the hybrid model. These metrics include accuracy (AC), sensitivity (SN), specificity (SP), F1-score (F1-S), and the Area Under Receiver Operating Characteristic Curve (AUC) [9], [10].

Therefore, the contribution in this paper to improving and solving some problems related to this field, which can be summarized as follows:

- We propose HDPF which consists of DBSCAN-based and resampling techniques (such as under-sampling, over-sampling, and hybrid) are used to solve the imbalance problem and eliminate the outliers.
- Analyzing and indexing a large amount of heart disease patient features, including elements that contain different categories in type and quantity, such as numbers, texts, etc. We use multiple datasets (two different large datasets).
- Preprocessing data in terms of deleting redundant data and treating missing data so that the classifier can work well.
- Extract the most important features using a simple genetic algorithm that can be relied upon in classifying data and reducing their numbers without compromising the data's accuracy so that we can reduce the time consumed.
- We use Density-based spatial clustering of applications with noise (DBSCAN) which is used to group closely features together (features with many nearby neighbors), marking as outliers features that lie alone in low-density regions.
- Can be built a hybrid classifier from supervised machine learning algorithms (such as LR, SVM, KNN, DT, RF) to classify existing data and predict new data for similar cases.
- Overfitting and underfitting problems are handled using 10-folds cross-validation.
- Performance analysis and comparison with state-of-the-art models.

The remainder of this manuscript is divided into five sections. Section II reviews previous work related to early diagnosis and health care for some diseases in general, and it focuses on heart diseases. Section III describes the data used and the challenges that must be overcome while dealing with this type of data. Section IV explains the different stages of the proposed framework and novel algorithms. Section V describes the experimental results of different test cases and the discussion section. Finally, section VI will provide conclusions and references.

II. RELATED WORKS

In this section, some related studies regarding recent modalities in healthcare and disease diagnostics will be reviewed. Researchers, academic scholars, and data scientists have undertaken various research initiatives in predicting and screening medical data for heart diseases. Multiple ML and data mining algorithms have been used in recent studies to carry out these predictions.

Therefore, these relevant works will be reviewed and compared with our proposed system. Most of the studies tend for analysis, and decision support systems are typically implemented using two various approaches. The first approach combines many features such as age, sex, chest, blood pressure, cholesterol, blood sugar, electrocardiographic results, heart rate, and several significant vessels colored by fluoroscopy, thalassemia, etc. The second approach reduces and restricts input patterns that can be easily measured [11], [12].

Desai *et al.* [13] utilized a novel classification model using a Backpropagation Neural Network (BPNN) and LR on the Cleveland heart disease dataset. Accuracies of 85.74% and 92.58% were recorded BPNN and LR respectively.

Padmanabhan *et al.* [12] proposed an approach of using Auto-Machine Learning (AutoML) in addition to the human expert system. The authors evaluated two cardiovascular disease datasets performance and compared the results to an AutoML library and human expert system. The accuracy and area under the curves for AutoML are significantly higher and better than those of the human expert system. Additionally, the time consumed by AutoML to produce these results is significantly less than the time consumed by the human expert system.

Islam *et al.* [14] proposed some superior data analysis techniques such as Naive Bayes (NB), LR, DT. In this case, LR provided the highest accuracy with 86.25%. Abhishek *et al.* [15] performed a heart disease forecast framework using the R programming language. The training and testing patterns are produced by dividing datasets into 70% and 30%, respectively. The test results showed that the NB classifier achieved a higher accuracy of 89%.

Rabbi *et al.* [16] conducted a comparative study on remarkable current classification models used in data mining such as SVM, KNN, and Artificial Neural Networks (ANN). Their test results showed that SVM achieved higher accuracy than both the KNN and ANN with 85% accuracy.

Dwivedi [17] performed a classification model using LR for predicting heart disease on the Cleveland dataset. They achieved 85% accuracy.

Abdeldjouad *et al.* [15] used a Genetic Fuzzy System-Logit Boost (GFS-LB), and Fuzzy Hybrid Genetic-Based Machine Learning (FH-GBML). The performance evaluation of these algorithms was implemented using WEKA [18] and KEEL tools. The highest accuracy of 80% was gained by majority voting.

Haq *et al.* [19] performed a classification model using LR with some preprocessing techniques for predicting heart disease on the Cleveland dataset. They achieved 89% accuracy.

Ali *et al.* [20] suggested an authority system based on stacked SVM to aid heart failure analysis. The primary SVM model was applied to exclude irrelevant features, while the second model was applied as a predictive model. They achieved 92% accuracy.

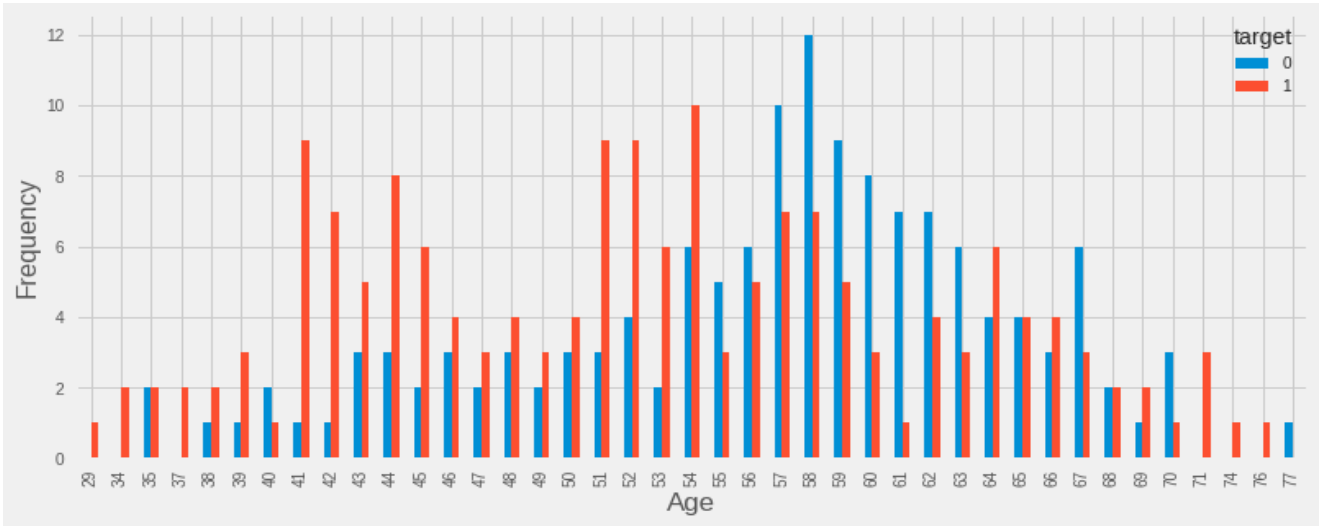

FIGURE 1. The distribution of the 'Target' feature against 'Age' feature.

Gupta *et al.* [21] performed a classification model using RF with some preprocessing techniques for predicting heart disease on the Cleveland dataset. They achieved 96.9% accuracy.

Fitriyani *et al.* [10] performed a classification model using XGBOOST + DBSCAN with some preprocessing techniques for predicting heart disease on the Statlog dataset and on the Cleveland dataset. They achieved an accuracy of 95% by using the Statlog dataset while the accuracy of 98% was achieved using the only Cleveland dataset. Table 1 represents the comparison between recent previous related works.

Amin *et al.* [23] proposed a hybrid technique with Naïve Bayes and Logistic Regression to predict cardiovascular disease. This research aims to identify significant features to improve the accuracy. They achieved 87.4% accuracy.

TABLE 1. Summary of recent previous related works.

Reference	Authors	Methods	ACC %
[17]	Dwivedi et al	LR	85
[19]	Haq et al.	LR	89
[22]	Saqlain et al.	SVM	81.19
[8]	Latha et al.	Voting between LR+MP+RF	85.4
[20]	Ali et al.	Stacked SVM	92
[21]	Gupta et al.	RF	96.9
[10]	Fitriyani et al.	XGBOOST + DBSCAN (Statlog dataset)	95
[23]	Amin et al.	NB + LR	87.4

III. GLOBAL CHALLENGES AND DESCRIPTIVE DATA ANALYSIS

A. GLOBAL CHALLENGES

Due to the nature of the dataset used, some challenges must be faced and overcome in the proposed model. These challenges are summarized as follows [10], [11]:

- Dataset contains wide variation and diversity features with high dimensionality. We use DBSCAN-based and resampling techniques to solve the imbalance problem and eliminate the outliers.
- Any diagnostic system requires high speed and accuracy to perform the tasks. We use a novel algorithm to extract the semantic features using simple genetic algorithm for reducing dimensionality without compromising the data's accuracy so that we can reduce the time consumed.
- There are redundant and missing data within the dataset being used. We use a novel preprocessing algorithm to solve this problem.

B. DATASET DESCRIPTION

We used the UCI database of cardiology. It contains four datasets that have been previously used by ML researchers. The ''target'' attribute indicates the appearance or nonexistence of heart disease in the patient [12]. This dataset contains 76 features. These features are smoking, body mass, physical activity, a healthy diet, cholesterol levels, blood pressure, fasting blood glucose, etc. These attributes are the same seven ideal measures that the American Heart Association has set to promote cardiovascular health and disease reduction [6]. The four databases contain redundant and sometimes missing data [30], [31]. We will reduce the number of investigated

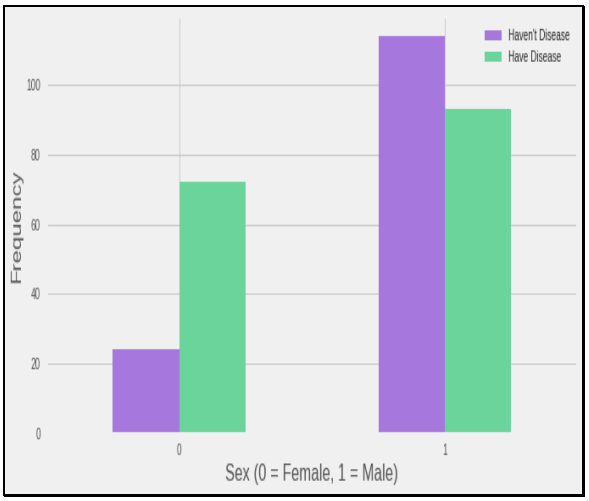

FIGURE 2. The distribution of cardiovascular patients by gender.

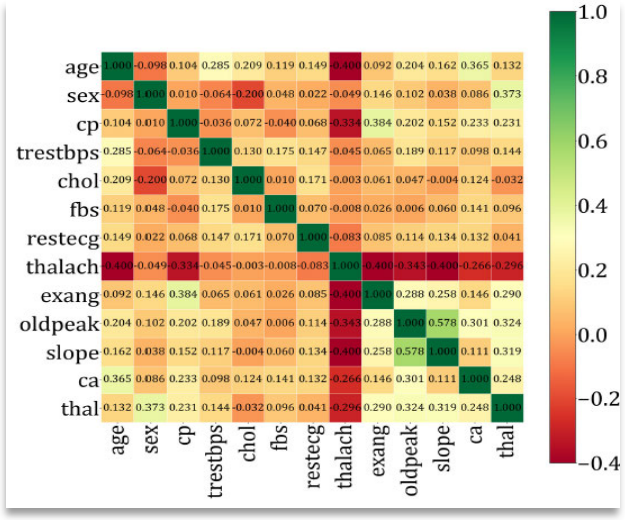

FIGURE 3. The correlation heatmap for the dataset.

attributes to 14. We will use algorithms that only select the best 14 of 76 attributes to minimize feature dimensionality. Data were collected from four different datasets: the Cleveland Foundation, Statlog, the Budapest Institute of Cardiology, the California Medical Center, and the Zurich Hospital. Table 2 shows the description of the 14 features used in our model.

C. DESCRIPTIVE DATA ANALYSIS

In this section, we will review a statistical and investigative analysis of the data used. The distribution of the target feature among the remaining features will also be studied. We found that the age group '55-60' occupied the distribution peak. Figure 1 shows the distribution of the 'target' feature against 'age' feature. Additionally, the distribution of cardiovascular patients by sex is shown in Figure 2. We found that the largest numbers of people suffering from heart diseases were in the range of '41-64' years old. Patients in the 20-30 age group are less likely to suffer from heart disease [13]–[15]. Using the 'describe ()' function from the Pandas library, we obtain various descriptive statistics that exclude NaN values. Several descriptive statistics are returned as the count, mean, standard deviation, minimum-maximum values, and data quantiles. As shown in Table 3, most values are generally categorized. Mean values tell us the average value of that feature. Using the Python Matplotlib library functions, we can explore the correlation between the attributes of the dataset by visualizing it as shown in the heat map in Figure 3. It is clear that the degree of correlation between the 'target' and the rest of the data variables is weak.

IV. HEART DISEASE PREDICTION FRAMEWORK (HDPF)

This section will propose an intelligent framework to diagnose heart disease using machine learning and a Simple Genetic Algorithm (SGA). The proposed framework aims to diagnose heart disease early and help doctors make the appropriate decision to reduce mortality. One of the main challenges is the wide variation and diversity of features in the data. Therefore, we will process the data to extract the features and derive new features for machine learning which are more accurate and faster. The proposed framework contains three different phases. The first phase cleans the data by deleting duplicates, imputing missing data, and normalization, called preprocessing. The second phase includes primary processing, such as extracting features and deriving additional features from the data based on SGA. SGA operates based on crossover and mutation to generate synthetic chromosomes from the original population or set of factors. These chromosomes that produce high fitness values remain, while the others drop out. The mutation step is conducted at the end, in which the global search is maximized, and the best value is found. Finally, the chromosome describes the picked feature.

Finally, the third phase applies a hybrid method of machine learning algorithms to classify data. Figure 5 shows the proposed framework. In algorithm 1, there are several steps. First, the performance vector is initialized using well-known performance metrics such as accuracy, AUC, precision, recall, and the F1-score. Next, the flowchart of HDPF has several steps, including Data Imputation and Partitioning, DBSCAN, SGA, Feature Extraction, Machine Learning (ML) Approach, and Performance Metric Evaluation as shown in figure 6. Finally, to make the dataset complete and reasonable for processing, data imputation is done to fill the missing values of the features with the new labels.

Second, we use algorithm 2 to perform data preprocessing such as data cleaning, data imputation, and feature normalization. Some well-known equations are used to perform feature normalization preprocessing. These equations are considered as follows:

$$\mu_j = \frac{1}{n}\sum_{x=1}^{n} x_j, x_j \in D \tag{1}$$

TABLE 2. Description of features used in the dataset [7].

ID	Code	Feature Name	Data type	Measurement	Domain Range	Missing Values
1	age	AGE	Continuous-Real	Age in years	29 - 79	NO
2	sex	SEX	Discrete Binary	Patient Gender	0, 1	NO
3	cp	CPT	Discrete Categorical	1 = typical angina; 2 = atypical angina. 3 = non-angina 4 = asymptomatic	1, 2, 3, 4	YES
4	trestbps	RBP	Continuous Real	mm Hg	94 – 200	NO
5	chol	CHOL	Continuous Real	mg/dl	126 – 564	NO
6	fbs	FBS	Discrete Binary	mg/dl	0, 1	YES
7	restecg	EGR	Discrete Categorical	0 = normal. 1 = ST-T wave abnormal. 2 = left ventricular	0, 1, 2	NO
8	thalach	MHR	Continuous Real	Numeric	71 – 202	NO
9	exang	EIG	Discrete Binary	Binary	0, 1	NO
10	oldpeak	ST depression	Continuous-Real	Numeric	0 – 6.2	NO
11	slope	The slope of the Peak Exercise	Discrete Categorical	1 = upsloping. 2 = flat; 3 = downsloping	1, 2, 3	NO
12	ca	NMV	Discrete Real	Colored by fluoroscopy	0 – 3	YES
13	thal	Exercise thallium scintigraphy	Discrete Categorical	3 – Normal, 6 – Fixed Defect, 7 –Reversible Defect	3, 6, 7	YES
14	Class	Target	Discrete Binary	0 = absence; 1 = presence	0 or 1	NO

TABLE 3. Statistical analysis of the dataset [8].

	age	sex	cp	treetops	Chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
Unique Values	41	2	4	50	152	2	3	91	2	40	3	4	3	2
mean	54.37	0.68	0.97	131.62	246.26	0.15	0.53	149.65	0.33	1.04	1.40	0.73	2.31	0.54
std	9.08	0.47	1.03	17.54	51.83	0.36	0.53	22.91	0.47	1.16	0.62	1.02	0.61	0.50
min	29.00	0.00	0.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	0.00	0.00	0.00	0.00
max	77.00	1.00	3.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	2.00	4.00	3.00	1.00

$$\sigma_j = \sqrt{\frac{1}{n}\sum_{x=1}^{n}(x_j - \mu_i)^2}, x_j \in D \tag{2}$$

$$x_i = \frac{x_i - \mu_i}{\sigma_j} \tag{3}$$

where μ = mean, σ = standard deviation, D = dataset, n = total number of values, x = single feature value. These data features are normalized using one unit mean and zero variance.

SGA is a scientific representation determined by the famous Charles Darwin's approach based on Biological pick[26]. Natural selection processes just the most qualified individuals over several periods. In machine learning, the use of SGA is to take the best amount of variables to produce a favorable treatment [29].

Preparing the perfect part of variables is an investment of combinatory and optimization. The advantage of this method over others is that it provides the most suitable assistance to emerge from the various helpful prior solutions. An evolutionary formula that promotes the option in time. The idea of SGA is to combine the multiple solutions along many periods after production to extract the most helpful genetics (variables) from each one.

We can determine several other uses of GA, such as hyper-tunning specification, find the maximum (or minutes) of a feature, or look for a correct neural network design (Neuroevolution), or among others [29]. To calculate fitness value (FV), we include an optional weight W for the selection probability. By default, W = 1 means that the candidate solutions' fitness fully determines the selection probability for six crossovers. If W is set to values smaller than 1, the

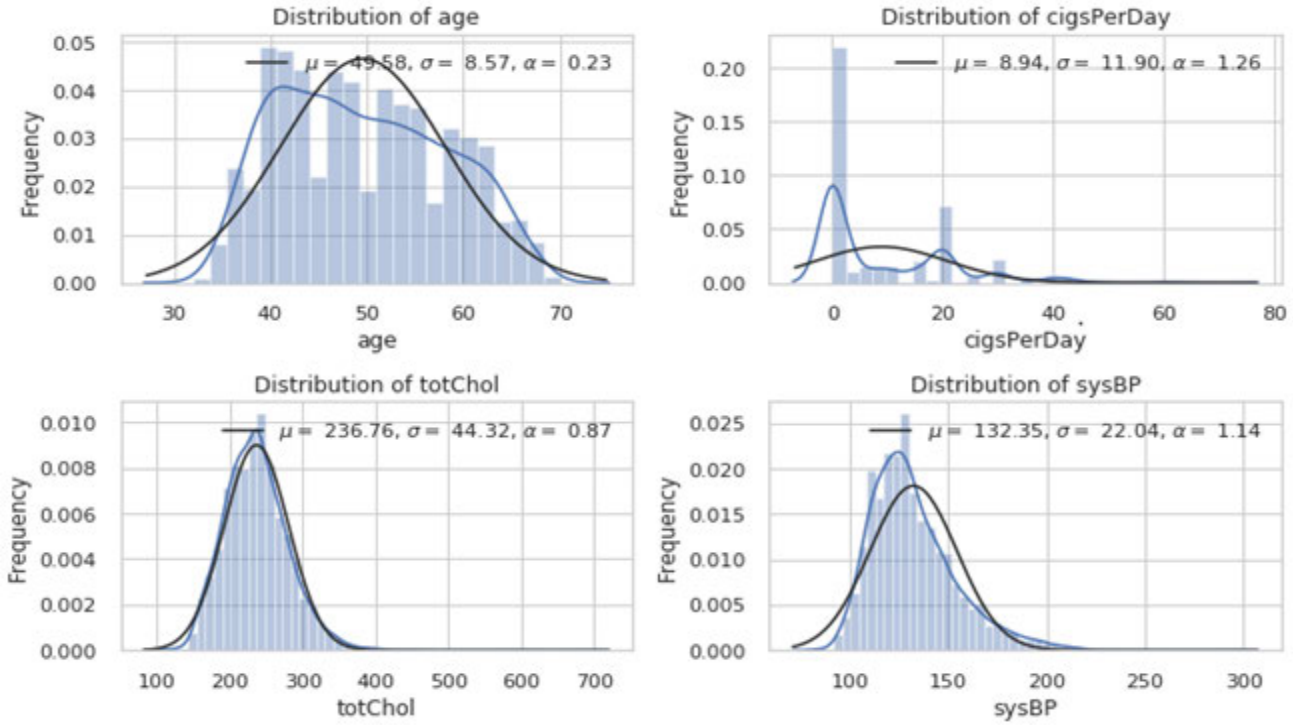

FIGURE 4. Statistical charts for several features of heart disease.

Algorithm 1: HDPF Pseudocode

```
Input: Heart Disease Dataset (D)
Output: Vector of Performance Metrics (PM)
Start Procedure
    PV ← Performance Vector {AC,Pre,Fm}
    D' ← Data_preprcoessing(D)
    Fx ← Feature Extraction (D')
    Optimal features ← Feature Selection (Fx)
    Foreach approach:{RF, DT, KNN, SVM, LR} do
        ML_method ← ML_algorithm (approach)
        For i=1:length(Optimal features) do
            (wAC,wAUC,wF1,wSen) ← ML_method(i)
            Performance ← wAC * AC + wAUC * AUC
              +wSen + sen)
        End For
    End Foreach
    Ensembled_algo ← highest performance of three
    ML_method
    Foreach Optimal features and Ensembled_algo do
        PV ← Ensembled_algo (Optimal_features)
    End Foreach
    Return PV
End Procedure
```

importance of the individual fitness decreases. If W = 0, the selection probability is independent of the fitness so that the chance of being chosen for a crossover would be equal for every candidate solution. Fitness value was calculated according to equation (5).

$$\boldsymbol{Fpro_i} = \frac{\boldsymbol{FV_i}}{\sum_{i=1}^{n=6} \boldsymbol{FV_i}} \tag{4}$$

$$\boldsymbol{FV_i} = (\boldsymbol{1} - \boldsymbol{W})\,(\boldsymbol{t} - \boldsymbol{1}) + \boldsymbol{x^i(t)} \tag{5}$$

Equation 4 is used for fitness probability estimation to a single gene type. Fpro ith is fitness probability. FV ith is fitness value. In Equation 5, the search space is denoted by $x^i(t)$, t represents time, and i mean feature level. The summation of cumulative fitness values should be equal to 1. Pick the maximum fitness value j and check if it satisfies the condition $csj < csk$ where csj is the cumulative sum, and csk is the newly formed subsequence set, as shown in Figure 7. Convergence is the state where we arrive at an optimal solution with leading fitness values. Fourth, the feature selection step is applied in Algorithm 3.

Third, SGA is applied to the data to obtain old and newly derived features. This step is considered feature extraction. Figure 6 shows the flowchart of optimal feature search. Figure 8 shows the flowchart of the SGA general structure. Next, principal component analysis (PCA) is applied to obtain qualitative label features (QLF) and quantitative numeric features (QNF). The optimal features are selected from algorithm 3 by maximizing the squared correlation coefficient summation between QLF, FX, and QNF, and FX. Finally, the classification step is obtained. We used the splitting holdout function to divide the data into training and validation datasets where factor = 0.2. We use 10-fold cross-validation to overcome

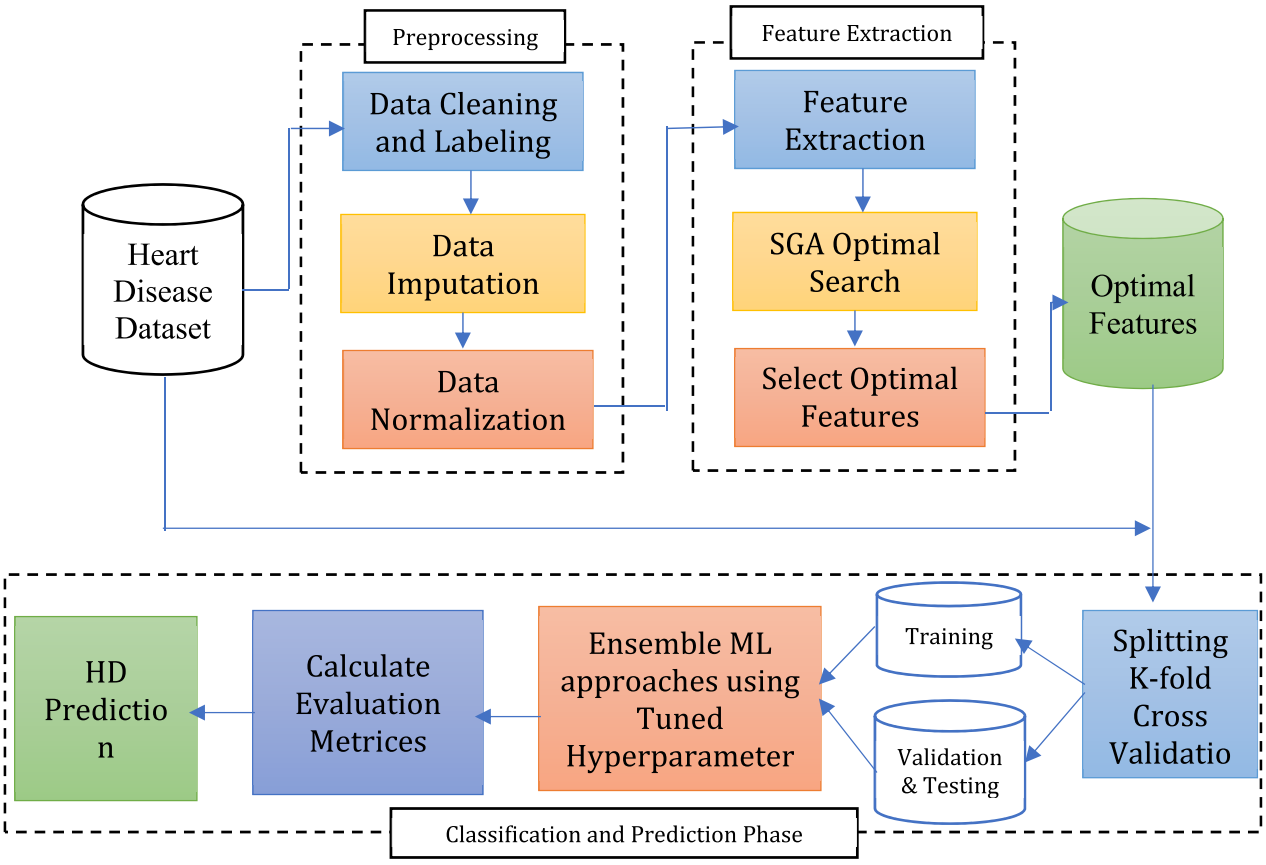

FIGURE 5. HDPF Framework.

overfitting problems. Several ML algorithms are applied, and the three algorithms' highest accuracy is chosen to perform the ensembled process.

V. EXPERIMENTAL RESULTS ANALYSIS AND DISCUSSION

In this section, the results of our various experiments will be explained and compared to relevant previous research. A heart disease dataset extracted from the UCI Machine Learning Repository was used and is described in Section III. All tests were conducted on Intel Core i7 2.90 GHz CPU and 8 GB RAM. We use Python as the programming language to develop different tasks.

A. RESULTS AND PERFORMANCE MEASURES

We used 10-folds cross-validation to avoid overfitting problems. Also, we analyzed and enumerated the model's performance during the learning phase. Finally, the dataset was divided into test and train sets. Dataset separated utilizing dimension 70:30, i.e., 70% from the data for training and 30% for testing the model, which is the standard dimension for partitioning datasets. The upside of this partitioning is that it provides sufficient information to prepare and test the proposed framework.

Moreover, it manages away from under-fitting if the training partition is smaller than the testing samples. Additionally, if the training partition is more distinguished than the testing partition, this can overfitting the framework. We used classification metrics such as sensitivity (SN), specificity (SP), accuracy (AC), and F1-score (F1-S) to measure the model's efficiency. The equations for those metrics can be listed as follows:

$$SN = \frac{TP}{TP + FN} \times 100\% \tag{6}$$

$$SP = \frac{TN}{TN + FP} \times 100\% \tag{7}$$

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \times 1\,00\% \tag{8}$$

$$F1 - S = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \times 100\% \tag{9}$$

where TP = true positive, FN = false negative, FP = false positive and TN = true negative.

Our study used HDPF and SGA to determine the optimal features for our recommended framework. The initial information about SGA factors is as follows: the initial population is set randomly at 100, the number of periods used is 100 with crossover and mutation probability of 0.5 and 0.001, respectively. The experimental outcomes exposed that the proposed framework achieved an accuracy of 98.18%. The accuracy obtained by the suggested framework using SGA

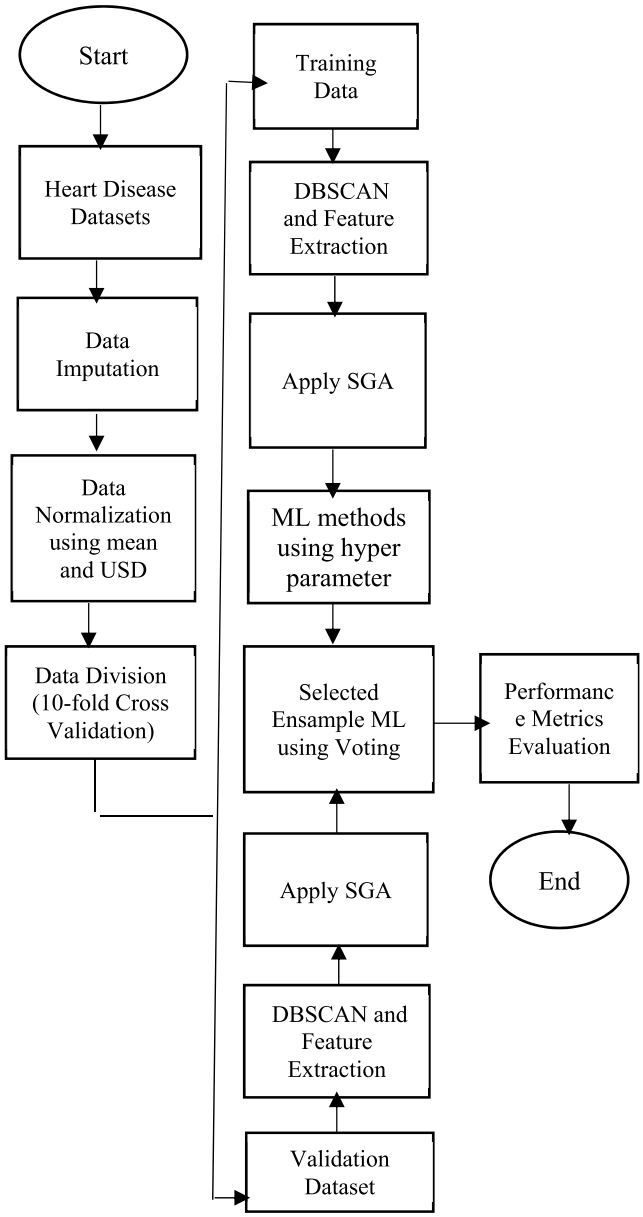

FIGURE 6. Flow chart of HDPF.

Algorithm 2: Preprocessing Pseudocode

Input: Heart Disease Dataset (D)
Output: Processed Data (D')

Start Procedure
  For i=1: length (D)
  D'(i) ← remove duplication
  //data imputation
  If i ∈ D and i is a categorical label value
    If D(i) == missing value
      D'(i) ← Majority value of that field
    Else
      Continue
    End If
  End If
  If *i ∈ D and i is a numerical label value*
    If D(i) == missing value
      D'(i) ← mean value of that field
    Else
      Continue
    End if
  End If
  //data normalization
  For j=1:length(D)
    $\mu_j = \frac{1}{n}\sum_{x=1}^{n} x_j,\ x_j \in D$
    $\sigma_j = \sqrt{\frac{1}{n}\sum_{x=1}^{n}(x_j - \mu_i)^2},\ x_j \in D$
    $x_i \leftarrow \frac{x_i - \mu_i}{\sigma_j}$
  End For
  Return D'
End Procedure

Algorithm 3: Feature Selection Pseudocode

Input: Extracted Feature (FX)
Output: Optimal Selected Feature (FC)

Start Procedure
  QLF, QNF ← PCA(FX)
  For i=1:length(QLF)
    C← correlation coefficient between QLF and Fx
    S ← correlation coefficient between QNF and Fx
  End For
  Optimal_features ← max($\sum_{QLF} C^2\,(QLF, Fx) + \sum_{QLF} S^2\,(QNF, Fx)$)
End Procedure

has improved by 3.18% compared to the accuracy performed by other related models. The recommended HDPF model also achieved 98% precision and 0.98 F1-Score. In addition, we have conducted experiments by using several supervised machine learning algorithms and different numbers of extracted features. As represented in Table 4, we found that DT and RF achieved the highest precision and accuracy than the other algorithms. They were also the least time-consuming to implement the processing.

Confusion matrices are drawn in Figure 8. From Figure 8(e), we find that the confusion matrix of the RF algorithm has achieved the largest total of true positive and true negative. Also has the lowest sum of the values false negative and false positive compared to the rest of the participated algorithms.

LR and SVM came second in achieving total TP and TN, as shown in Figures 9(a) and 9(c). But SVM algorithm was the best out of all in error type II, where achieved zero TN.

As for DT and KNN algorithms, they were the lowest achieved values in total TP and TN. And at the same time, they had the highest value in errors type I and II, as shown in Figures 9(b) and 9(d).

Figure 10 represents the ROC curve for the members participating in the Hybrid classification technique. The biggest

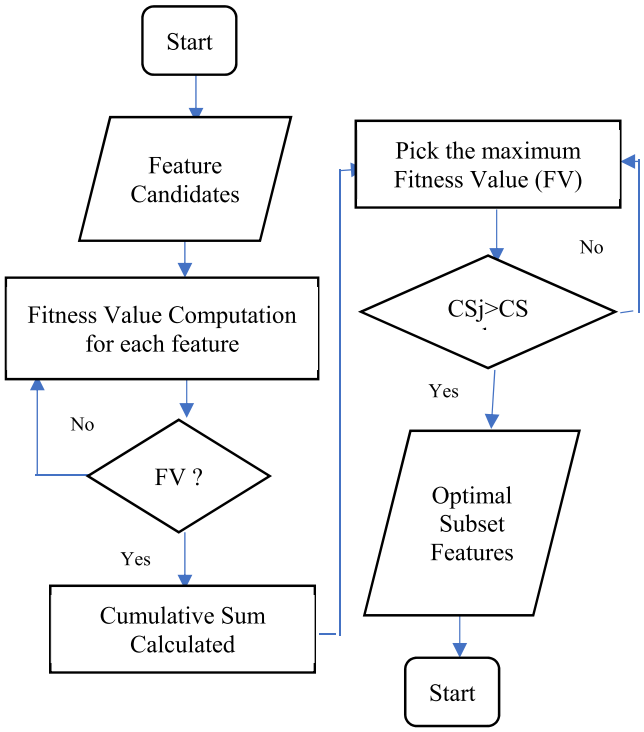

FIGURE 7. Flow chart of SGA optimal search.

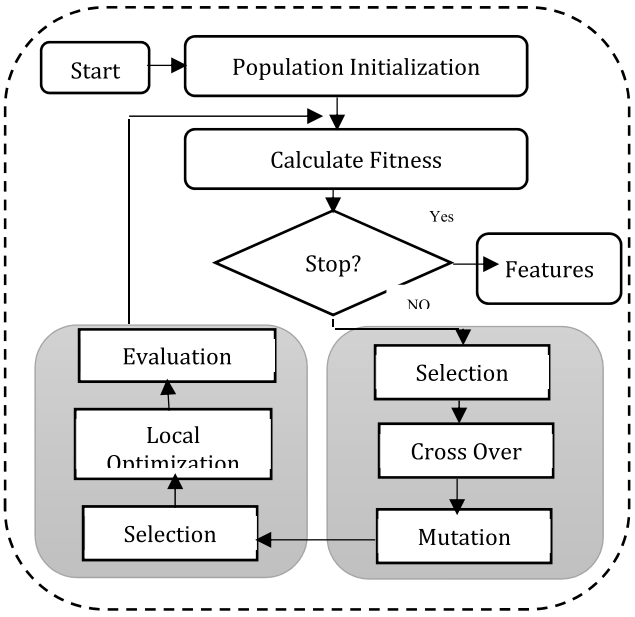

FIGURE 8. Flow chart of SGA general structure.

AUC was achieved under the curve of both algorithms, the Random Forest and Decision Tree. At the same time, the KNN algorithm achieved the minor area under the curve.

B. DISCUSSION

Features extraction played a crucial role in determining evaluation metrics for the supervised machine learning algorithms participating in this experiment. We notice that by using feature sets (11, 17, 23, 28) and applying them to the same algorithm, we find that the evaluation metrics have been affected by them.

TABLE 4. Performance Results and Time consumption Using Various Supervised Machine Learning Algorithms.

Method	# features	AC %	SN %	SP %	F1-S %	Time msec.
LR	11	88.52	89.28	87.78	87.71	43
	17	91.89	92.85	90.90	91.22	59
	23	88.52	92.85	84.84	88.13	82
	28	88.52	93.14	93.93	88.79	106
SVM	11	85.24	85.71	84.84	84.21	33
	17	86.88	89.28	84.50	86.20	42
	23	83.60	0.75	90.91	80.76	55
	28	91.80	100	90.95	91.80	67
KNN	11	77.04	82.14	72.7	76.66	44
	17	77.04	67.85	84.84	73.07	61
	23	81.96	71.80	90.91	78.43	84
	28	81.32	71.42	90.78	80.92	102
DT	11	90.16	92.85	87.97	89.65	40
	17	91.17	94.77	90.12	90.20	56
	23	93.39	92.98	91.91	92.13	77
	28	93.19	92.67	91.90	92.29	98
RF	11	91.80	89.28	93.93	90.90	39
	17	93.81	89.38	93.56	90.14	51
	23	92.80	96.42	87.78	91.52	71
	28	94.44	96.28	97.96	95.59	85

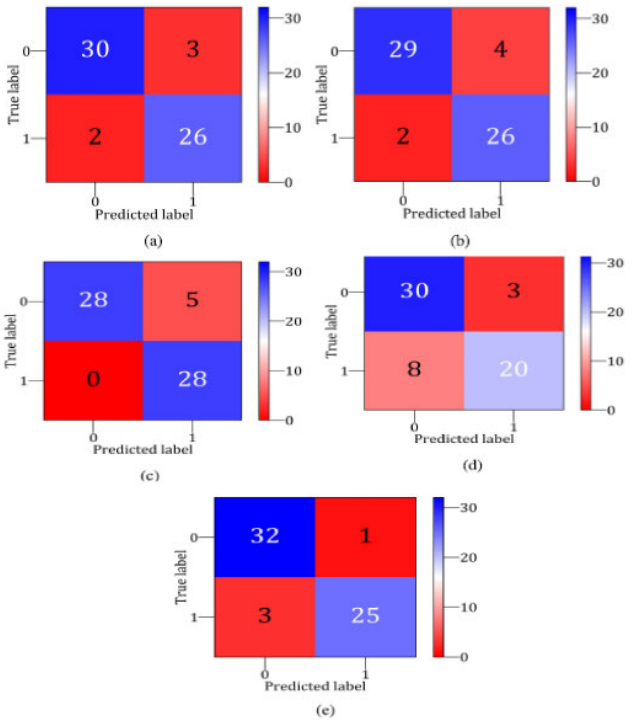

FIGURE 9. Confusion matrices (a) LR (b) KNN (c) SVM (d) DT and (e) RF.

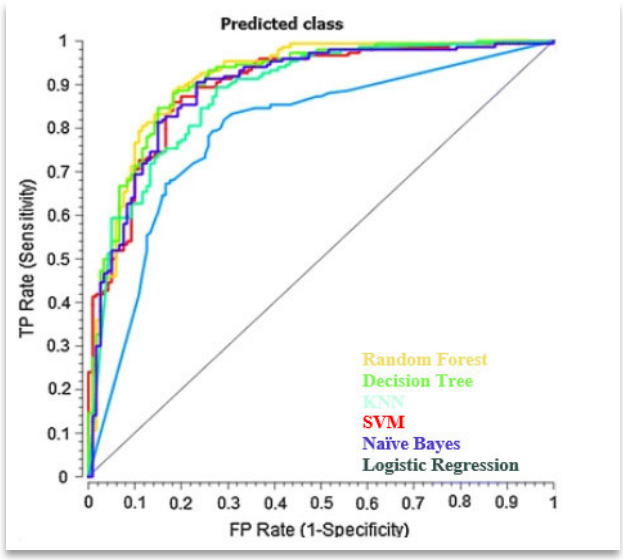

FIGURE 10. ROC (Sensitivity and Specificity) of hybrid classification techniques.

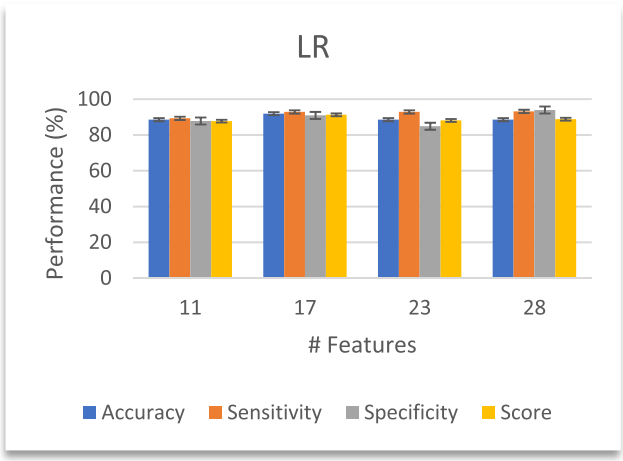

FIGURE 11. LR performance chart.

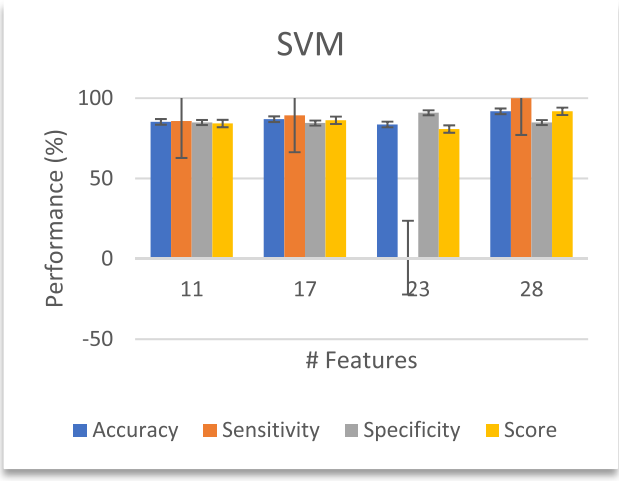

FIGURE 12. SVM performance chart.

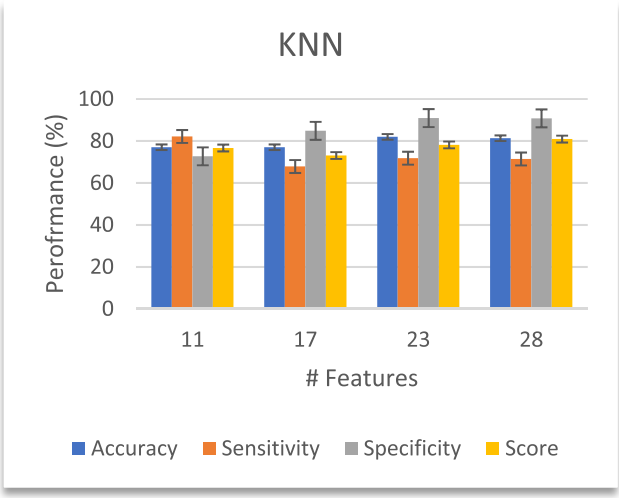

FIGURE 13. KNN performance chart.

According to Table 4, several features are extracted, and then supervised machine learning algorithms are applied. In addition, we found DT and RF achieved higher accuracy than other algorithms. Also, the average time consumption for DT and RF is less than different algorithms. So, the proposed ensemble algorithm is a hybrid model between DT and RF.

The performances of different ML methods change according to the number of features used. For example, Table 3 and Figures 10, 11, 12, 13, and 14 show that RF and DT achieved higher accuracy than the other algorithms. Therefore, we applied DT to training and validation datasets. Then, the majority voting ensemble technique was used to the result with RF to achieve high accuracy (98.18%).

Several experiments were conducted using multiple machine learning algorithms with SGA. First, several features with different levels were extracted (feature sets 11, 17, 23, and 28). Then, we found that the evaluation metrics have been affected by them. For example, Table 4 and figures from 11 to 15 showed that feature set 28 achieved high or equal accuracy with the previous feature set. For most cases, the more extracted features, the more precision we have. But, according to time consumption, feature sets 23 and 28 consumed more time.

C. COMPARISON BETWEEN PROPOSED FRAMEWORK AND PREVIOUS RELATED WORKS

The comparative analysis presented in Table 5 reveals that there is a significant difference in the performance of the proposed HDPF and other models. Visualizing these results through Figure 16, the proposed framework achieved the highest accuracy than the related works of 98.18%. While, Fitriyani *et al.* [10], Gupta *et al.* [21], and Ali *et al.* [20] achieved an accuracy of 95%, 93.4%, and 92%, respectively.

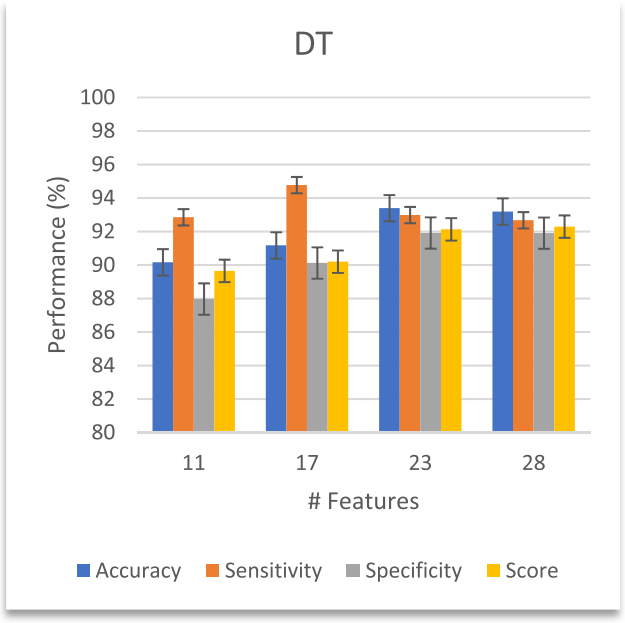

FIGURE 14. DT performance chart.

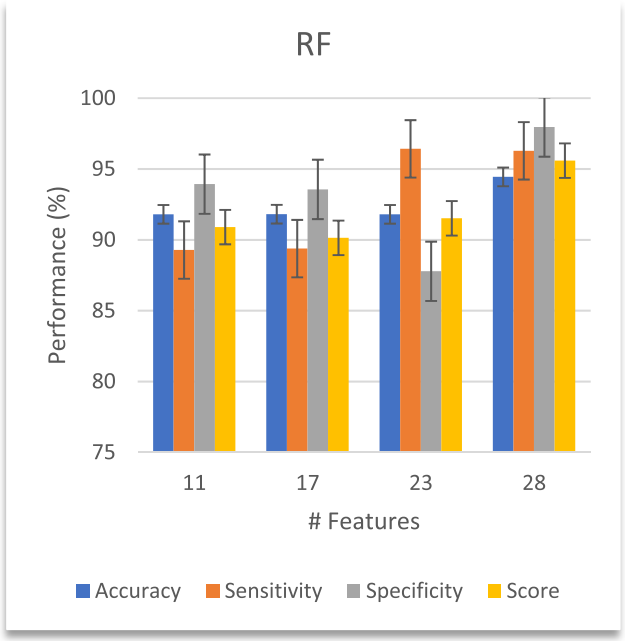

FIGURE 15. RF performance chart.

TABLE 5. Performance Comparison Between The Proposed Framework And Other Related Works.

Reference	Authors	Methods	AC %
[17]	Dwivedi et al	LR	85
[19]	Haq et al.	LR	89
[22]	Saqlain et al.	SVM	81.19
[8]	Latha et al.	Voting between LR+MP+RF	85.4
[20]	Ali et al.	Stacked SVM	92
[21]	Gupta et al.	RF	93.4
[10]	Fitriyani et al.	XGBOOST + DBSCAN (Statlog)	95
[23]	Amin et al.	NB + LR	87.4
	Proposed HDPF	SGA + Voting between LR+SVM+KNN+RF+ DT	98.18

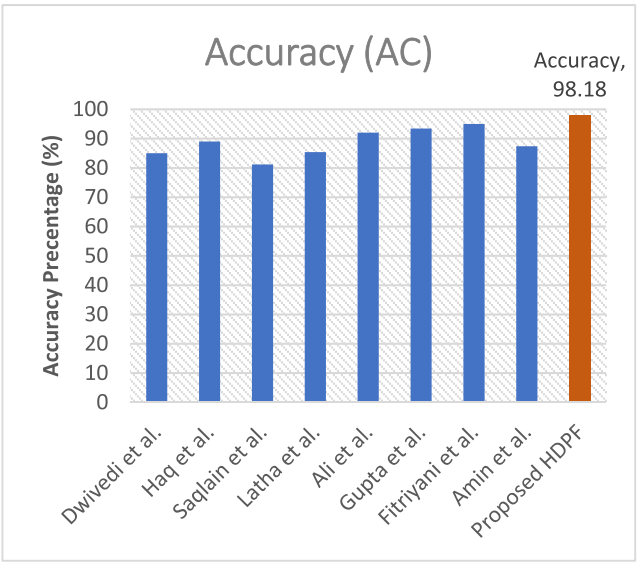

FIGURE 16. Comparison between proposed HDPF and other related works.

D. HEART DISEASE DECISION SUPPORT SYSTEM TO TEST THE PREDICTION SYSTEM

We designed and developed the proposed HDPF into Decision Support System (DSS) to diagnose the heart disease status effectively and efficiently. The DSS was developed using PHP version 7.2 scripting language and MYSQL version 8.0 database. Figure 17 shows the general structure of the DSS model. In DSS, the patient uses a web application through the local webserver (WAMP server) to enter diagnosis data such as Age, Sex, CP, thal,.. etc. Then, the proposed model was processed the input data using the proposed algorithms and hybrid ensembled machine learning to predict heart disease status. Figure 18 shows the result of DSS.

VI. CONCLUSION

This paper introduces hybrid classifiers using an ensembled model with a majority voting technique to improve prediction accuracy. Furthermore, a proposed preprocessing

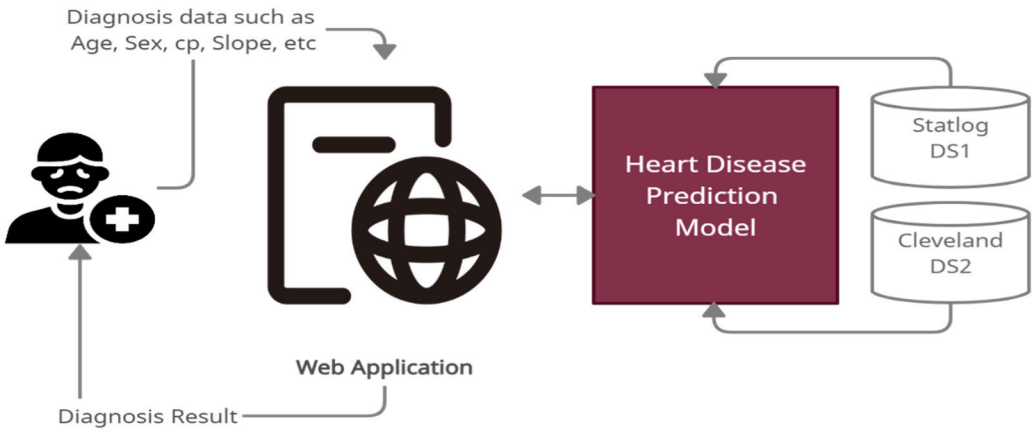

FIGURE 17. General structure of heart disease DSS.

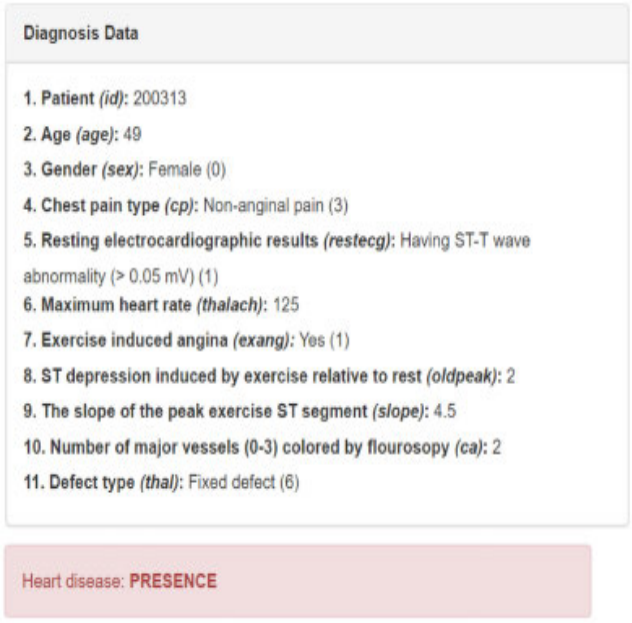

FIGURE 18. Result of DSS.

technique and feature selection based on a genetic algorithm is suggested to enhance prediction performance and overall time consumption. Experiments were performed on a dataset for cardiovascular patients from the UCI Machine Learning Repository. The study results indicated that the proposed ensemble classifier model achieved a classification accuracy of 98.18% through a comparative analytical approach. In comparison, the average performance of each machine learning algorithm gained 88%, 85%, 80%, 92%, and 93% for LR, SVM, KNN, DT, and RF, respectively. For future research, you can predict health status in real-time based on health-based streaming data as Twitter heart disease streaming data. In this paper, you will develop the proposed system using Twitter Streaming API, Apache Kafka, Apache Spark, and various machine learning models. Also, we can use a semantic ontology algorithm as in the published paper [32] to extract semantic features to enhance accuracy and reduce overall processing time. Since we have done the first stage of the system work in this paper to get the best machine learning model, a real-time online prediction pipeline will be attached as a second stage in the development work.

REFERENCES

[1] M. Heron, "Deaths: Leading causes for 2010," *Natl. Vital Stat. Rep.*, vol. 62, no. 6, pp. 1–96, 2013.
[2] *Cardiovascular Diseases (CVDs)*. Accessed: May 9, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[3] S. Chauhan and B. T. Aeri, "The rising incidence of cardiovascular diseases in India: Assessing its economic impact," *J. Preventive Cardiol.*, vol. 4, no. 5, pp. 735–740, 2015.
[4] G. A. Roth, G. A. Mensah, and C. O. Johnson, "Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study," *J. Amer. College Cardiol.*, vol. 76, no. 25, pp. 2982–3021, Dec. 2020.
[5] K. Uyar and A. Ilhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," *Proc. Comput. Sci.*, vol. 120, pp. 588–593, Jan. 2017.
[6] S. Shilaskar and A. Ghatol, "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4146–4153, Aug. 2013.
[7] S. Ismaeel, A. Miri, and D. Chourishi, "Using the extreme learning machine (ELM) technique for heart disease diagnosis," in *Proc. IEEE Canada Int. Humanitarian Technol. Conf. (IHTC)*, May 2015, pp. 1–3.
[8] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informat. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100203.
[9] G. S. Handelman, H. K. Kok, R. V. Chandra, A. H. Razavi, S. Huang, M. Brooks, M. J. Lee, and H. Asadi, "Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods," *Amer. J. Roentgenol.*, vol. 212, no. 1, pp. 38–43, Jan. 2019.
[10] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 133034–133050, 2020.
[11] K. Saxena and U. Banodha, "A fuzzy logic based cardiovascular disease risk level prediction system in correlation to diabetes and smoking," in *Data Management, Analytics and Innovation* (Advances in Intelligent Systems and Computing), vol. 1042. Singapore: Springer, 2020, pp. 29–40.
[12] M. Padmanabhan, P. Yuan, G. Chada, and H. V. Nguyen, "Physician-friendly machine learning: A case study with cardiovascular disease risk prediction," *J. Clin. Med.*, vol. 8, no. 7, p. 1050, Jul. 2019.

[13] S. D. Desai, S. Giraddi, P. Narayankar, N. R. Pudakalakatti, and S. Sulegaon, "Back-propagation neural network versus logistic regression in heart disease classification," in *Advanced Computing and Communication Technologies* (Advances in Intelligent Systems and Computing), vol. 702. Singapore: Springer, 2019, pp. 133–144.
[14] S. Islam, N. Jahan, and M. E. Khatun, "Cardiovascular disease forecast using machine learning paradigms," in *Proc. 4th Int. Conf. Comput. Methodol. Commun. (ICCMC)*, Mar. 2020, pp. 487–490.
[15] F. Z. Abdeldjouad, M. Brahami, and N. Matta, "A hybrid approach for heart disease diagnosis and prediction using machine learning techniques," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Lecture Notes in Computer Science), vol. 12157. Cham, Switzerland: Springer, 2020, pp. 299–306.
[16] M. F. Rabbi, M. P. Uddin, M. A. Ali, and M. F. Kibria, "Performance evaluation of data mining classification techniques for heart disease prediction," *Amer. J. Eng. Res.*, vol. 7, no. 2, pp. 278–283, 2018.
[17] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Comput. Appl.*, vol. 29, no. 10, pp. 685–693, 2018.
[18] *Weka 3–Data Mining With Open Source Machine Learning Software in Java*. Accessed: May 9, 2021. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/
[19] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Inf. Syst.*, vol. 2018, Dec. 2018, Art. no. 3860146.
[20] L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour, and S. A. C. Bukhari, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure," *IEEE Access*, vol. 7, pp. 54007–54014, 2019.
[21] A. Gupta, R. Kumar, H. Singh Arora, and B. Raman, "MIFH: A machine intelligence framework for heart disease diagnosis," *IEEE Access*, vol. 8, pp. 14659–14674, 2020.
[22] S. M. Saqlain, M. Sher, F. A. Shah, I. Khan, M. U. Ashraf, M. Awais, and A. Ghani, "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," *Knowl. Inf. Syst.*, vol. 58, no. 1, pp. 139–167, Jan. 2019.
[23] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics Inform.*, vol. 36, pp. 82–93, Mar. 2019.
[24] M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, and C. S. Pattichis, "Assessment of the risk factors of coronary heart events based on data mining with decision trees," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 3, pp. 559–566, May 2010.
[25] Y. Mok, Y. Sang, S. H. Ballew, C. M. Rebholz, W. D. Rosamond, G. Heiss, A. R. Folsom, J. Coresh, and K. Matsushita, "American Heart Association's Life's simple 7 at middle age and prognosis after myocardial infarction in later life," *J. Amer. Heart Assoc.*, vol. 7, no. 4, Feb. 2018, Art. no. e007658.
[26] *Braunwald's Heart Disease: A Textbook of Cardiov–9780323462990*. Accessed: May 9, 2021. [Online]. Available: https://www.us.elsevierhealth.com/braunwalds-heart-disease-a-textbook-of-cardiovascular-medicine-single-volume-9780323462990.html
[27] L. Ali, S. U. Khan, N. A. Golilarz, I. Yakubu, and I. Qasim, "A feature-driven decision support system for heart failure prediction based on statistical model and Gaussian naive Bayes," *Comput. Math. Methods Med.*, vol. 2019, Nov. 2019, Art. no. 6314328.
[28] J. K. Kim and S. Kang, "Neural network-based coronary heart disease risk prediction using feature correlation analysis," *J. Healthcare Eng.*, vol. 2017, Sep. 2017, Art. no. 2780501.
[29] B. P. Doppala, D. Bhattacharyya, M. Chakkravarthy, and T.-H. Kim, "A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset," *Distrib. Parallel Databases*, vol. 2021, pp. 1–20, Mar. 2021.
[30] *Statlog (Heart) Data Set*. Accessed: May 9, 2021. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/statlog+(heart)
[31] *Heart Disease Data Set*. Accessed: May 9, 2021. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease
[32] M. M. El-Gayar, N. E. Mekky, A. Atwan, and H. Soliman, "Enhanced search engine using proposed framework and ranking algorithm based on semantic relations," *IEEE Access*, vol. 7, pp. 139337–139349, 2019.

SARRIA E. A. ASHRI received the B.Sc. and M.Sc. degrees in industrial engineering from the Faculty of Engineering, Mansoura University, Mansoura, Egypt, in 1986 and 2017, respectively. In 2012, she has been awarded a Postgraduate Diploma in information technology from the Faculty of Computers and Information, Mansoura University. Since 2017, she has been working as a Researcher with the Faculty of Computers and Information. Her research interests include industrial engineering and management, renewable and sustainable energy engineering, cloud computing, data science, big data analytics, and the Internet of Things (IoT).

M. M. EL-GAYAR received the B.Sc., M.Sc. and Ph.D. degrees in information technology from Mansoura University, Mansoura, Egypt, in 2010, 2013, and 2020, respectively. Since 2020, he has been working as a Lecturer with the Faculty of Computer and Information Science. His research interests include semantic web, algorithms, big data, deep learning, computer vision, image processing, big data, and pattern recognition.

EMAN M. EL-DAYDAMONY received the B.S., M.S., and Ph.D. degrees in electrical communications from the Faculty of Engineering, Mansoura University, Egypt, in 1998, 2003, and 2008, respectively. From 1998 to 2002, she worked as a Teaching Assistant. From 2003 to 2007, she worked as a Teacher, and from 2008 to 2010, she worked as a Lecturer. From 2010 to 2017, she worked as a Lecturer with the Faculty of Computers and Information Sciences (FCIS), Mansoura University. Since 2018, she has been working as an Associate Professor at FCIS. Her current research interests include computer vision, pattern recognition, medical image processing, biomarker discovery, and bioinformatics.