IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 7, 2020, accepted July 12, 2020, date of publication July 20, 2020, date of current version July 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010511

HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System

NORMA LATIF FITRIYANI[1], MUHAMMAD SYAFRUDIN[1], GANJAR ALFIAN[2], (Member, IEEE), AND JONGTAE RHEE[1]

[1]Department of Industrial and Systems Engineering, Dongguk University, Seoul 04620, South Korea
[2]Industrial AI Research Center, Nano Information Technology Academy, Dongguk University, Seoul 04626, South Korea

Corresponding authors: Muhammad Syafrudin (udin@dongguk.edu) and Jongtae Rhee (jtrhee@dongguk.edu)

This work was supported by the Dongguk University Research Fund, in 2019, under Grant S-2019-G0041-00035.

ABSTRACT Heart disease, one of the major causes of mortality worldwide, can be mitigated by early heart disease diagnosis. A clinical decision support system (CDSS) can be used to diagnose the subjects' heart disease status earlier. This study proposes an effective heart disease prediction model (HDPM) for a CDSS which consists of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to detect and eliminate the outliers, a hybrid Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE-ENN) to balance the training data distribution and XGBoost to predict heart disease. Two publicly available datasets (Statlog and Cleveland) were used to build the model and compare the results with those of other models (naive bayes (NB), logistic regression (LR), multilayer perceptron (MLP), support vector machine (SVM), decision tree (DT), and random forest (RF)) and of previous study results. The results revealed that the proposed model outperformed other models and previous study results by achieving accuracies of 95.90% and 98.40% for Statlog and Cleveland datasets, respectively. In addition, we designed and developed the prototype of the Heart Disease CDSS (HDCDSS) to help doctors/clinicians diagnose the patients'/subjects' heart disease status based on their current condition. Therefore, early treatment could be conducted to prevent the deaths caused by late heart disease diagnosis.

INDEX TERMS Heart disease, disease prediction model, clinical decision support system, outlier data, imbalanced data, machine learning.

I. INTRODUCTION

Heart disease is a cardiovascular disease (CVD) that remains the number one cause of death globally and contributes to approximately 30% of all global deaths [1]. If unmitigated, the total number of deaths globally is projected to increase to around 22 million in 2030. The American Heart Association reported that nearly half of American adults are affected by CVDs, equating to nearly 121.5 million adults [2]. In Korea, heart disease is among the top three leading causes of death and contributed to nearly 45% of total deaths in 2018 [3]. Heart disease is a condition when plaque on arterial walls can block the flow of blood and cause a heart attack or stroke. Several risk factors that can lead to heart disease include unhealthy diet, physical inactivity, and excessive use of tobacco and alcohol. These risk factors can be minimized by practicing good daily lifestyle such as salt reduction in the diet, consuming fruits and vegetables, doing regular physical activity, and discontinuing use of tobacco and alcohol which eventually could help to reduce the risk of heart disease [4]. The early heart disease identification of high-risk individuals and the improved diagnosis using a prediction model have generally been recommended to reduce the fatality rate and improve the decision-making for further prevention and treatment [5]–[7]. A prediction model that is implemented in the clinical decision support system (CDSS) can be used to help clinicians assess the risk of heart disease and provide appropriate treatments to manage the risk further [8]. In addition, numerous studies have also reported that the implementation of CDSS can improve preventive care, clinical decision making and decision quality [9]–[12].

Machine learning-based clinical decision making have recently been applied in healthcare area. Previous studies have shown that machine learning algorithms (MLAs) such as chaos firefly algorithm [13], backpropagation neural network (BPNN) [14], multilayer perceptron (MLP) [15],

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

logistic regression (LR) [16], support vector machine (SVM) [17], and random forest (RF) [18] have been successfully used to help as decision making tools for heart disease prediction based on individual data. Several studies have also revealed the advantage of a hybrid model which achieved good performance in predicting heart disease such as majority voting of naïve bayes (NB), bayes net (BN), RF, and MLP [19], two stacked SVMs [20], and RF with a linear model [21]. However, in the machine learning field, outlier and imbalance data may arise and impact on the performance of the prediction model. Previous studies have reported that by incorporating Density-Based Spatial Clustering of Applications with Noise (DBSCAN)-based to detect and eliminate the outlier data [22]–[24], and by balancing the distribution of data using a hybrid Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE-ENN) [25]–[28], the prediction models' performances were significantly enhanced.

To the best of our knowledge, no study has investigated a heart disease prediction model (HDPM) by utilizing DBSCAN, SMOTE-ENN and XGBoost machine learning. Therefore, we propose an effective HDPM for a CDSS which consists of DBSCAN-based to detect and eliminate the outliers, SMOTE-ENN to balance the training data distribution and XGBoost to predict heart disease. Our challenge is to detect and remove the outlier data and to balance the distribution of the training dataset to improve the performance of the HDPM. Two publicly available datasets (Statlog [29] and Cleveland [30]) were used to build the model and to evaluate their performance compared with that of other models (NB, LR, MLP, SVM, decision tree (DT), and RF) and of previous study results. In addition, we ensured the applicability of the proposed model by designing and implementing the model into a Heart Disease CDSS (HDCDSS) to diagnose the subjects based on their current condition. The developed HDCDSS is expected to help clinicians diagnose the patients effectively and efficiently and thereby improving heart disease clinical decision making. Therefore, early treatment could be conducted to prevent the deaths caused by late heart disease diagnosis. Contributions of our study can be summarized as follows.

- *Improving accuracy of heart disease prediction model*. We proposed HDPM by integrating DBSCAN outlier detection, SMOTE-ENN, and XGBoost to improve prediction accuracy. The HDPM learned from two public datasets and the trained model was utilized to predict the subjects' heart disease status based on their current condition.
- *Performance analysis and comparison with state-of-the arts models*. The proposed HDPM was evaluated with other classification models and compared with the results from previous studies. In addition, we presented the statistical evaluation to confirm the significant of our model as compared to other models.
- *Real case system development*. We designed and developed the prototype of the system to show the feasibility and applicability of our proposed model for real-world case study. It is expected that the developed system can be used as a practical guideline for the healthcare practitioners.

The remainder of this study is organized as follows. Section II summarized the literature review. Section III presents the proposed HDPM including datasets description, overall design, and modules of the proposed model as well as performance evaluation metrics. Section IV discusses the performance evaluation of proposed model, including the statistical test and comparison with previous studies. Section V presents the practical applications of the proposed model in the real case scenario. Finally, the concluding remarks and future research directions are presented in Section VI.

II. LITERATURE REVIEW

Several studies have reported the development of heart disease diagnosis based on machine learning models with the aim of providing an HDPM with enhanced performance. Two publicly available heart disease datasets, namely Statlog and Cleveland, have been widely used to compare the performance of prediction models among researchers. For Statlog dataset, a heart disease clinical decision support system based on chaos firefly algorithm and rough sets-based attribute reduction (CFARS-AR) was developed by Long *et al.* (2015) [13]. The rough sets were used to reduce the number of attributes while the chaos firefly algorithm was used to classify the disease. The developed model was then compared with other models such as NB, SVM and ANN. The results revealed that the proposed model achieved the highest performance among all the models with accuracy, sensitivity, and specificity of 88.3%, 84.9%, and 93.3%, respectively. The combination of rough sets-based attributes selection and BPNN (RS-BPNN) was proposed by Nahato *et al.* (2015) [14]. With the selected attributes, the proposed RS-BPNN achieved accuracy of up to 90.4%. Dwivedi (2018) [31] compared six machine learning models (ANN, SVM, LR, k-nearest neighbor (kNN), classification tree and NB) with various performance metrics. The results showed that LR performed better than the other models by achieving up to 85%, 89%, 81%, and 85 for the accuracy, sensitivity, specificity, and precision, respectively. Amin *et al.* (2019) [32] performed comparison analysis by identifying significant attributes and applying machine learning models (k-NN, DT, NB, LR, SVM, Neural Network (NN) and a hybrid (voting with NB and LR)). The experiment results revealed that the hybrid model (voting with NB and LR) with selected attributes achieved the highest accuracy (87.41%).

Cleveland heart disease dataset has been widely used by researchers to generate predictive models. Verma *et al.* (2016) [15] developed a hybrid prediction model based on correlation feature subset (CFS), particle swam optimization (PSO), K-means clustering and MLP. The results showed that the proposed hybrid model achieved accuracy of up to 90.28%. Haq *et al.* (2018) [16] performed a comparative study on a hybrid model based on various

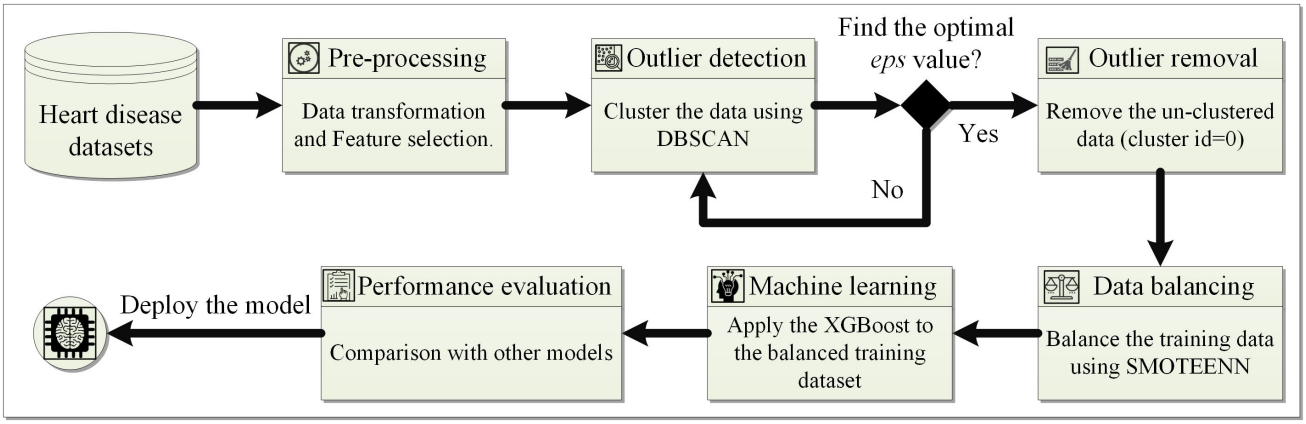

FIGURE 1. The proposed Heart Disease Prediction Model (HDPM) for the Heart Disease Clinical Decision Support System (HDCDSS).

feature selection techniques (relief, minimal-redundancy-maximal-relevance (mRMR), least absolute shrinkage and selection operator (LASSO)) and machine learning models (LR, kNN, ANN, SVM, DT, NB, and RF). Their study revealed that the features reduction affects the performance of the models. The study concluded that a combination of Relief-based feature selection and LR-based machine learning algorithm (MLA) provides higher accuracy (up to 89%) as compared with other combinations used in the study. Saqlain *et al.* (2019) [17] proposed a technique based on mean Fisher score feature selection algorithm (MFSFSA) and SVM classification model. The selected features are based on the higher Fisher score than the mean score. Then, SVM used the selected feature subset to learn and calculate the MCC through a validation process. The study revealed that the combination of FSFSA and SVM generates accuracy, sensitivity, and specificity of up to 81.19%, 72.92%, and 88.68%, respectively. Latha and Jeeva (2019) [19] proposed a hybrid model with majority voting of NB, BN, RF, and MLP. The proposed model achieved an accuracy of up to 85.48%.

Ali *et al.* (2019) [20] proposed two stacked SVMs to improve the diagnosis process. The first SVM was used to remove the non-relevant features and the second to predict heart disease. The results revealed that the proposed model achieved better performance than other models and previous study results. Mohan *et al.* (2019) [21] introduced a hybrid RF with a linear model (HRFLM) to enhance the performance of the HDPM. They found that the proposed method achieved accuracy, precision, sensitivity, f-measure and specificity of up to 88.4%, 90.1%, 92.8%, 90%, and 82.6%, respectively. Recently, Gupta *et al.* (2020) [18] developed a machine intelligence framework consisting of factor analysis of mixed data (FAMD) and RF-based MLA. The FAMD was used to find the relevant features and the RF to predict the disease. The experimental results showed that the proposed method outperformed other models and previous study results by achieving the accuracy, sensitivity, and specificity of up to 93.44%, 89.28%, and 96.96%, respectively.

None of the aforementioned previous studies have applied outlier detection and data balancing method to improve the accuracy of classification model, especially for the case of heart disease datasets. Thus, in this study we used outlier detection and data balancing methods to improve the model performance. In addition, the XGBoost classifier is then used to learn and generate the prediction model.We expect that our proposed model will achieve higher performance than that of state-of-the-art models and previous study results. Finally, we also design and develop the HDCDSS to help doctors/clinicians diagnose the patients'/subjects' heart disease status based on their current condition. Thus, early treatment could be conducted to prevent the risks further.

III. MATERIALS AND METHODS

The proposed HDPM was developed to provide high performance prediction in the presence or absence of heart disease given the current condition of the subjects. The flow-chart in Figure 1 shows how the proposed HDPM is developed. First, the heart disease datasets are collected. Second, the data pre-processing for data transformation and feature selection are conducted. Third, the DBSCAN-based outlier detection method is applied to find the outlier data given the optimal parameter. Fourth, the detected outlier data are then removed from the training dataset. Fifth, the data balancing based on SMOTE-ENN method is used to balance the training dataset. Sixth, the XGBoost-based MLA is used to learn from the training dataset and generate the HDPM. Finally, the performance metrics are presented to evaluate the performance of the proposed model and the generated HDPM is then implemented within the CDSS. In our study, we utilized 10-fold cross-validation method to avoid the overfitting. Cross-validation allows the models to learn from different sets of training data by repeated sampling; hence maximizing the data used for validation and possibly, helping to prevent from overfitting. Previous study has demonstrated that 10-fold cross-validation can be used to maintain the bias-variance trade-off which eventually provide the generalized model and protect against overfitting [33], [34].

TABLE 1. The detailed dataset attributes description and distribution (mean and standard deviation (STD)) for dataset I (Statlog).

Attributes					Present (Positive)	Absent (Negative)
No.	Symbol	Description	Type	Data Range	Mean ± STD	Mean ± STD
1	*age*	Subject age in years	Numeric	[29, 77]	56.59 ± 8.12	52.71 ± 9.51
2	*sex*	Subject gender	Binary	0 = female, 1 = male	-	-
3	*cp*	Chest pain type	Nominal	1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic	-	-
4	*trestbps*	Resting blood pressure in mmHg	Numeric	[94, 200]	134.44 ± 19.1	128.87 ± 16.46
5	*chol*	Serum cholesterol in mg/dl	Numeric	[126, 564]	256.47 ± 47.97	244.21 ± 54.02
6	*fbs*	Fasting blood sugar with value > 120 mg/dl	Binary	0 = false, 1 = true	-	-
7	*restecg*	Resting electrocardiographic result	Nominal	0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy	-	-
8	*thalach*	Maximum heart rate	Numeric	[71, 202]	138.86 ± 23.13	158.33 ± 19.28
9	*exang*	Exercise induced angine	Binary	0 = no, 1 = yes	-	-
10	*oldpeak*	ST depression induced by exercise relative to rest	Numeric	[0, 6.2]	1.58 ± 1.28	0.62 ± 0.8
11	*slope*	Slope of the peak exercise ST segment	Nominal	1 = up-sloping, 2 = flat, 3 = down-sloping	-	-
12	*ca*	Number of major vessels (0-3) colored by flourosopy	Nominal	0 – 3	-	-
13	*thal*	Defect type	Nominal	3 = normal, 6 = fixed defect, 7 = reversable defect	-	-

The detailed steps, including datasets and modules descriptions, and the performances metrics are presented in the following subsections. In addition, the performance of the proposed model with the state-of-the-art models is evaluated and the results are presented in the results and discussion section. Finally, we ensure the applicability of the proposed model by embedding the HDPM into the HDCDSS to diagnose the subjects' heart disease status based on their current condition.

A. HEART DISEASE DATASET

We used two heart disease datasets (Statlog and Cleveland; termed datasets I and II, respectively) to investigate how heart disease can be identified by applying the machine learning model. The proposed model is then applied to those two datasets and with the expectation of providing a general and robust HDPM.

The University of California Irvine (UCI) Repository Statlog Heart Disease database website presents dataset I to investigate heart disease [29]. The original dataset consists of 270 subjects, 13 attributes and one output class (120 and 150 subjects are labelled with the presence (positive class) and absence (negative class) of heart disease, respectively). There are no missing values in dataset I. A detailed attributes description (including data type and range) and distribution (mean and standard deviation (STD)) for dataset I are given in Table 1.

Dr. Robert Detrano, M.D., provided dataset II (Cleveland Heart Disease dataset) to investigate heart disease that was collected from the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation in California, United States [30]. The original dataset comprises 303 subjects and 79 raw attributes, although only 13 attributes are used, and one attribute as an output class. We removed 6 subjects' data due to missing values and used the remaining 297 data in the pre-processing stage. The original class value is a multi-class variable with the value range from 0 to 4. The 0 value is used to represent the absence of heart disease while the values from 1 to 4 are used to represent the presence of heart disease with its stage condition. In this study, we followed previous studies [16]–[21, [32] in converting the class value from a multi-class variable to a binary-class variable. The final class variable is set to 0 if heart disease is not present in the subject and to 1 for all the subjects who have been diagnosed as having heart disease. We pre-processed the data by applying the previous rule to the records. Finally, after data pre-processing, the final dataset II consists of 297 subjects with 137 and 160 subjects being labelled with the presence (positive class) and absence (negative class) of heart disease, respectively. A detailed attributes description (including data type and range) and distribution (mean and STD) for dataset II is given in Table 2.

For both datasets, the absence and presence of heart disease are treated as negative (0) and positive (1), respectively. The correlation between attributes can affect the performance of the machine learning model. Data correlation by utilizing Pearson's Correlation Coefficient (PCC) can be used as a calculation tool to determine the relationship between attributes. PCC varies from -1 to $+1$, with a positive and a negative value indicating a highly positive and highly negative correlation between the variables, respectively, and a value close to zero indicating a low correlation between them. The heatmap correlation between attributes for datasets I and II are given in Figure 2(a) and 2(b), respectively. The *gray* color

TABLE 2. The detailed dataset attributes description and distribution (mean and standard deviation (STD)) for dataset II (Cleveland).

Attributes					Present (Positive)	Absent (Negative)
No.	Symbol	Description	Type	Data Range	Mean ± STD	Mean ± STD
1	*age*	Subject age in years	Numeric	[29, 77]	56.76 ± 7.9	52.64 ± 9.55
2	*sex*	Subject gender	Binary	0 = female, 1 = male	-	-
3	*cp*	Chest pain type	Nominal	1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic	-	-
4	*trestbps*	Resting blood pressure in mmHg	Numeric	[94, 200]	134.64 ± 18.9	129.18 ± 16.37
5	*chol*	Serum cholesterol in mg/dl	Numeric	[126, 564]	251.85 ± 49.68	243.49 ± 53.76
6	*fbs*	Fasting blood sugar with value > 120 mg/dl	Binary	0 = false, 1 = true	-	-
7	*restecg*	Resting electrocardiographic result	Nominal	0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy	-	-
8	*thalach*	Maximum heart rate	Numeric	[71, 202]	139.11 ± 22.71	158.58 ± 19.04
9	*exang*	Exercise induced angine	Binary	0 = no, 1 = yes	-	-
10	*oldpeak*	ST depression induced by exercise relative to rest	Numeric	[0, 6.2]	1.59 ± 1.31	0.6 ± 0.79
11	*slope*	Slope of the peak exercise ST segment	Nominal	1 = up-sloping, 2 = flat, 3 = down-sloping	-	-
12	*ca*	Number of major vessels (0-3) colored by flourosopy	Nominal	0 – 3	-	-
13	*thal*	Defect type	Nominal	3 = normal, 6 = fixed defect, 7 = reversable defect	-	-

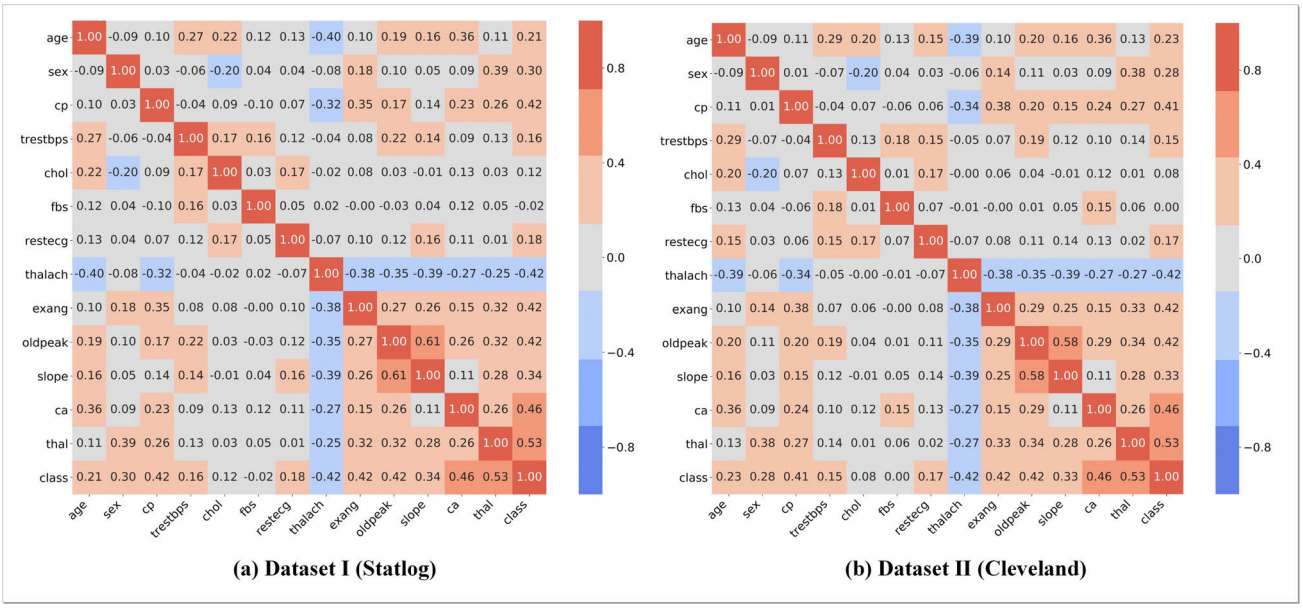

FIGURE 2. Heatmap of attributes correlation for (a) dataset I (Statlog) and (b) dataset II (Cleveland).

indicates that the correlation is close to 0, while the *red* and *blue* colors indicate that the correlation between variables is close to +1 and −1, respectively. The attributes *chol* and *fbs* are seen to have a correlation that is close to 0 toward the attribute *class*, which suggests that both only have a small or even no correlation with the attribute *class*. Thus, we could possibly remove these features to improve the performance of our proposed model.

In addition, we applied attribute selection by using the Information Gain (IG) method [35] in Weka V3.8 [36] to select the most important attribute to improve the model performance for the two datasets [37], [38]. Figure 3(a) and 3(b) show the attribute significant score based on the IG method for datasets I and II, respectively. In this case, both datasets have the same lowest attributes scores (*chol*, *trestbps*, and *fbs*), which we therefore removed from both datasets, and used the remaining attributes (*age, sex, cp, restecg, thalach, exang, oldpeak, slope, ca,* and *thal*) for further analysis. We expect that by using the two datasets and the selected attributes, our proposed model will be

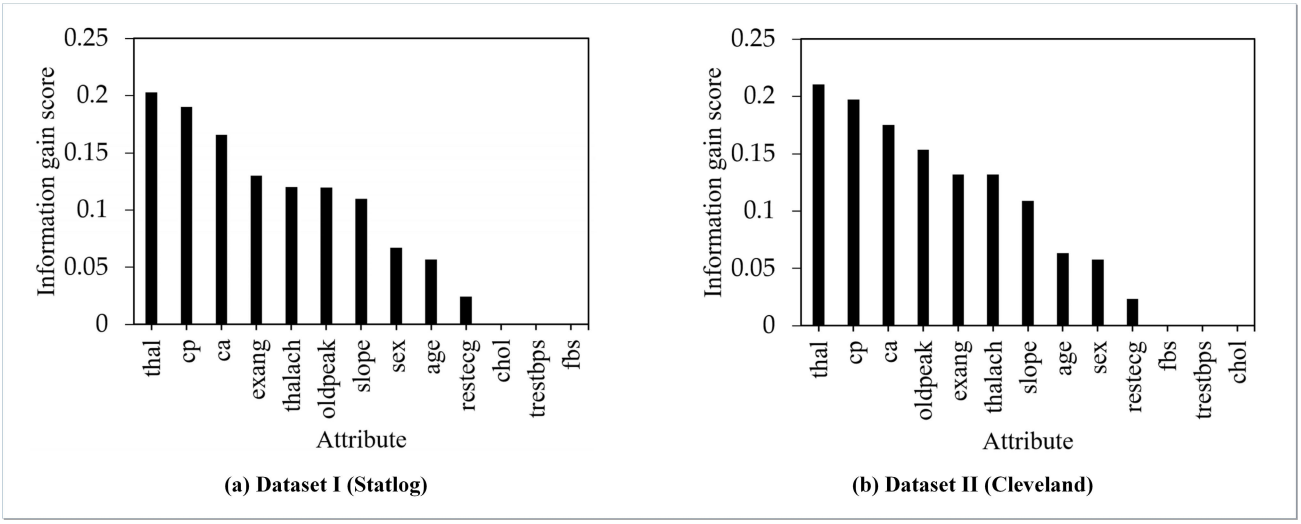

FIGURE 3. Attribute significance score provided by the Information Gain (IG) method for (a) dataset I (Statlog) and (b) dataset II (Cleveland).

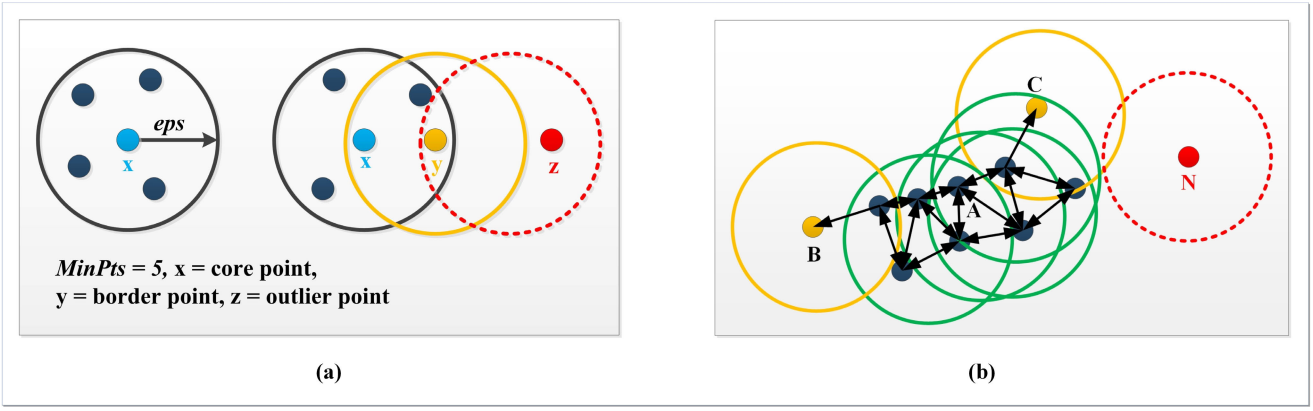

FIGURE 4. An illustration of (a) *eps*, core, border and outlier point and (b) DBSCAN cluster model with *MinPts* = 5.

sufficiently robust for predicting heart disease with high performance.

B. DBSCAN-BASED OUTLIER DATA DETECTION AND REMOVAL

In this study, we utilized DBSCAN [39] to cluster and detect the outliers from both training datasets. The goal of DBSCAN is to find the dense regions which can be identified by the number of objects that are close to a specific point (core point) and the points that are outside the regions are treated as outliers. In general, two parameters need to be determined for DBSCAN: epsilon (*eps*) and minimum points (*MinPts*). The *eps* is defined as the neighborhood radius around a point of x (ε-*neighborhood*) while the *MinPts* is defined as the minimum number of neighboring data points within the *eps*. There are three points that can be used to determine the normal and outlier data are core point, border point, and outlier point. A ''core point'' x is marked as any point that has a number of neighboring data points either greater than or equal to *MinPts*. The ''border point'' y is defined as the number of neighboring data points is less than *MinPts*, but y belongs to the neighboring core data point of x. Finally, the ''outlier point'' z is marked as a point z is neither a core point nor a border point. Figure 4(a) illustrates *eps*, core x, border y, and outlier z point using $MinPts = 5$. As can be seen in Figure 4(b), the point B and C are border point, A is a core point, and N is a noise point. Arrows indicate direct density reachability. Point B and C are density connected, because both are density reachable from point A. N is not density reachable and do not belong to any cluster (with $MinPts = 5$), and thus considered to be a noise point or outlier. First, the algorithm checks the specific point (any point) to be considered as a core point or not. The core point is if at least *MinPts* points are within the *eps* of it. The border points are the points that can be reached from core point (within distance *eps* from core point). Next, the core and border points are becoming cluster and marked as visited points by the algorithm. Finally, the algorithm keeps iterating to check other unvisited point

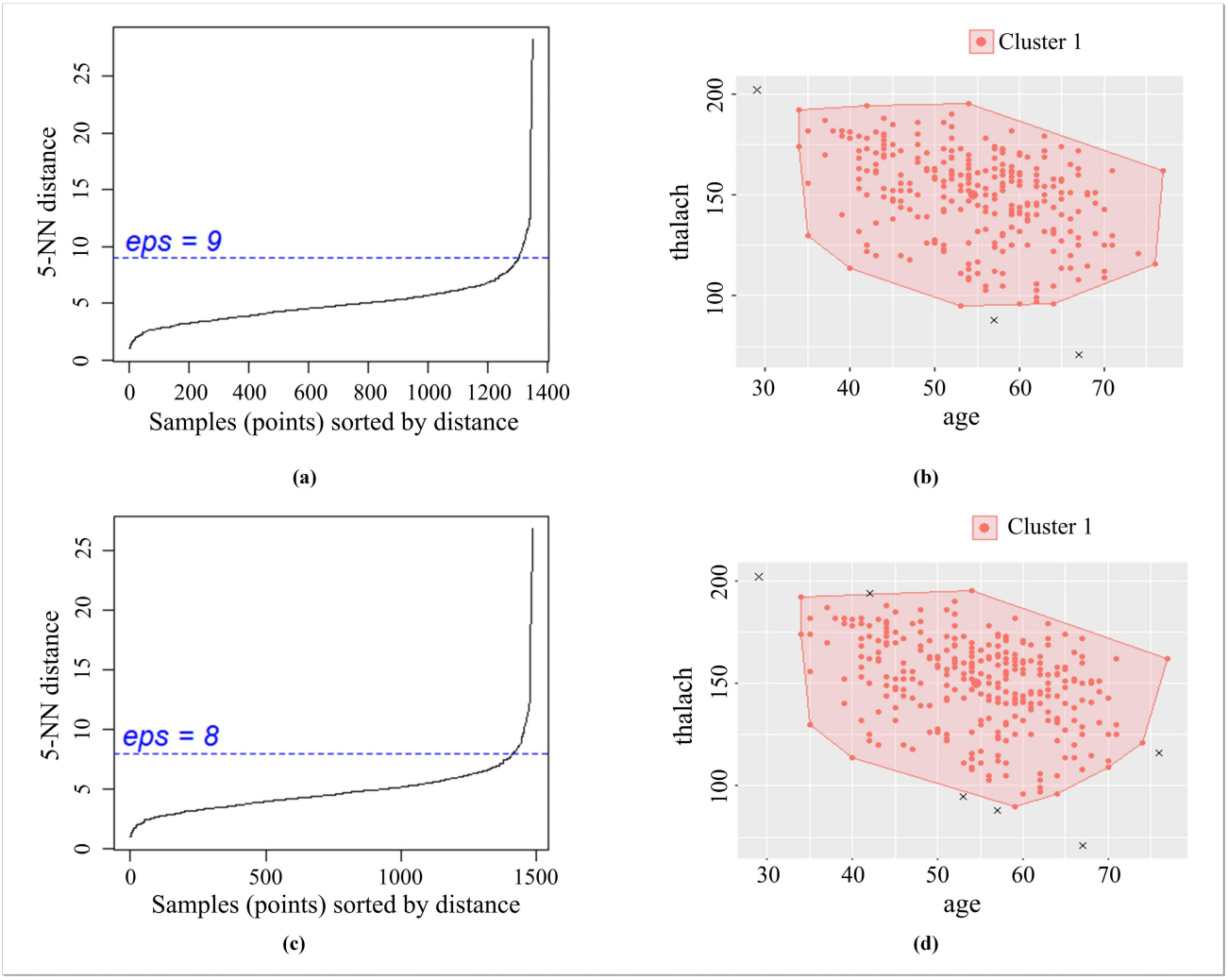

FIGURE 5. Optimal *eps* value using 5-NN and DBSCAN outlier detection result for datasets I (Statlog) (a), (b) and II (Cleveland) (c), (d), respectively.

(to be considered as core point) to find the unvisited border points. The points that are not belonging to the clusters are considered as outlier. The detailed pseudocode for DBSCAN is presented in Algorithm 1.

The optimal *eps* value is calculated by averaging the distance of every point to its kNN. The value of k corresponds to the *MinPts* value, which is defined by the user. In this study, we followed previous studies [40]–[43] to utilize 5-nearest neighbors (5-NN) to find the optimal *eps* value. Most of the previous studies utilized $MinPts = 5$ and optimized their *eps* value based on *MinPts*. Finally, according to Ester *et al.* (1996) [39], the *eps* can be obtained by presenting k-dist graph. First, k-distances are visualized as a k-dist graph and shown in ascending order to find the "*knee*" value where a sharp change appears beside the k-distance curve for the optimal *eps* value estimation. We implemented the calculation of kNNs and DBSCAN in R programming V3.5.1 and used R packages such as fpc V2.2-2 and DBSCAN V1.1-3.

Figure 5(a) and (c) show the sorted 5-NN distribution graph and optimal *eps* value for datasets I and II, respectively. We found that the "knee" appears at around the distance of 9 and 8 for datasets I and II, respectively. Furthermore, we applied the DBSCAN method by using $MinPts = 5$, $eps = 9$ and $MinPts = 5$, $eps = 8$ for datasets I and II, respectively. Figure 5(b) and (d) show the results of DBSCAN implementation for datasets I and II visualized in two-dimensional graphs. The results showed that in both datasets, the DBSCAN clustered the data into a single cluster as cluster 1 and the un-clustered data (with x symbol) are treated as outliers (see Figure 5(b) and (d)). The optimal parameters and the final outlier data for both datasets are presented in Table 3. Finally, we removed all the detected outlier data in each training dataset and used the remaining normal data for further analysis. In addition, we performed experimental analysis to find the impact of outlier removal on the performance of the model. Figure 6 shows the impact of outlier data elimination based on DBSCAN as compared to original data. Outlier removal based on DBSCAN significantly improved the model accuracy for all datasets, from accuracy 80.74%, 80.03% to 85.41%, 85.26% for dataset I, and II,

Algorithm 1 DBSCAN Pseudocode

```
Input: dataset, D; minimum point, minPts; radius, eps
Output: clustered C and un-clustered data UC
for each sample point SP in dataset D do
  if SP is not visited then
    mark SP as visited
    neigbrPts ← samples points in ε-neighborhood of SP
    if sizeof(neigbrPts) < minPts then
      mark SP as UC
    end
    else
      add SP to new cluster C
      for each sample point SP' in neigbrPts do
        if SP' is not visited then
          mark SP' as visited
          neigbrPts' ← samples points in
          ε-neighborhood of SP'
          if sizeof(neigbrPts') ≥ minPts then
            neigbrPts ← neigbrPts + neigbrPts'
          end
        end
        if SP' is not a member of any cluster then
          add SP' to cluster C
        end
      end
    end
  end
end
```

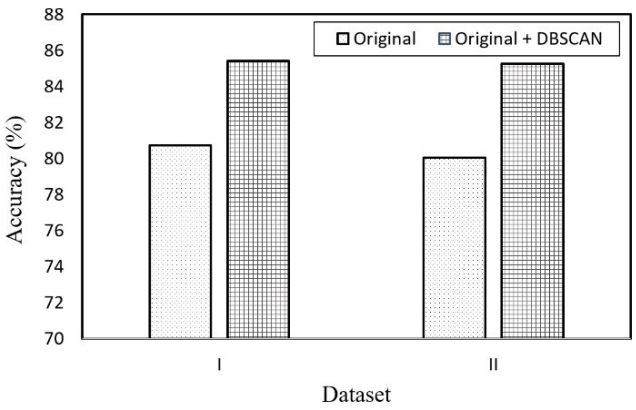

FIGURE 6. Impact of DBSCAN-based outlier elimination on model accuracy.

respectively, with average improvement as much as 4.95%. Furthermore, previous studies [22]–[24] also revealed that by removing outlier data, it has improved the performance accuracy.

C. SMOTE-ENN-BASED DATA BALANCING

Data sampling or data balancing is a common method comprised of three subcategories, over-sampling, under-sampling, and hybrid method, and is used in machine learning

TABLE 3. The parameters and result of DBSCAN-based outlier detection.

Dataset	*MinPts*	*eps*	# Outlier Data
Dataset I (Statlog)	5	9	3
Dataset II (Cleveland)	5	8	6

TABLE 4. SMOTE-ENN data balancing results.

Dataset	Before SMOTE-ENN		After SMOTE-ENN	
	Minority class (%)	Majority class (%)	Minority class (%)	Majority class (%)
I	44.19	55.81	50.79	49.21
II	46.05	53.95	49.5	50.5

to deal with imbalanced data. Figure 7 illustrates the three subcategories of data balancing methods. The over-sampling method balances the training data by generating data samples for the minority class while the under-sampling achieves that goal by eliminating the data samples in the majority class. Meanwhile, the hybrid method achieves the balanced data by combining the over-sampling and under-sampling methods.

We used a hybrid SMOTE-ENN [25] method to balance the imbalance heart disease training datasets. In general, SMOTE is used to over-sample the minority class until the training dataset is balanced, then the Edited Nearest Neighbor (ENN) is used to eliminate the unwanted overlapping samples between two classes while maintaining the balanced distributions. The pseudocode of SMOTE-ENN is explained in Algorithm 2. Previous studies have shown that the combination of SMOTE and ENN (SMOTE-ENN) provides better performances than that of either alone [25], [26]. For all datasets, the minority and majority classes are the subjects who were diagnosed with the presence (positive class) and absence (negative class) of heart disease, respectively. The original percentage of minority class over the total number of subjects for datasets I and II are 44.19% and 46.05%, respectively. The SMOTE technique was applied to increase the number of minority class by randomly generating new samples from the NNs of the minority class sample. Then the ENN was used to remove the unwanted overlapping samples. After SMOTE-ENN implementation, the total number of minority class increases, and the updated percentage of minority class for datasets I and II becomes more balanced, at 50.79% and 49.5%, respectively. We utilized Python V3.6.5 and the Imbalanced-learn python library V0.4.3 [44] to implement SMOTE-ENN, producing evenly balanced class distributions (see Table 4).

The SMOTE-ENN ensures that when creating the new artificial samples and eliminating the overlapped samples, it will follow the distribution pattern from the original samples. Figure 8 shows the data distribution of attributes "*age*" and "*thalach*" before and after SMOTE-ENN implementation for all training datasets. For each dataset, the distribution of attributes "*age*" and "*thalach*" follow the normal distribution pattern. The SMOTE-ENN implementation keeps the original data distribution pattern of dataset I, as shown

Algorithm 2 SMOTE-ENN Pseudocode

Input Data, D;
Output Balanced data, BD

foreach data point in minority class mp of data D **do**
    Compute the k-nearest neighbor Kmp_i
    Generate new synthetic data point $mp_{new} = mp_i + (\hat{mp}_i - mp_i) + \delta$
    Add the mp_{new} to D with mp_i class
end for
foreach data point p in data D **do**
    if $p_i class$ <> majority class of k-nearest neighbors **then**
        Remove p_i from D
    end if
end for
return BD

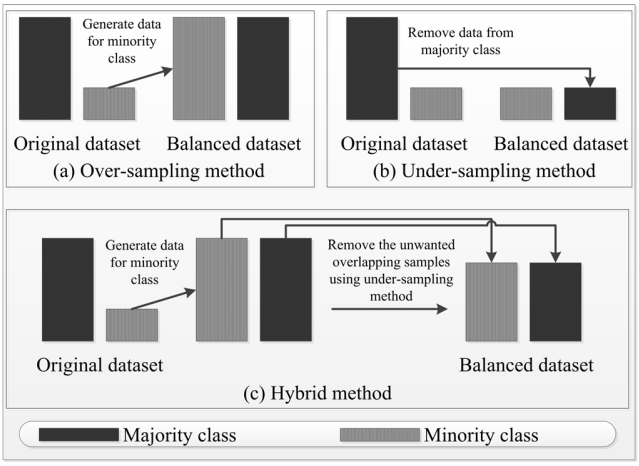

FIGURE 7. Impact of DBSCAN-based outlier elimination on model accuracy.

in Figure 8(b), such that the updated dataset I retains a similar pattern of data distribution (normal distribution). Dataset II exhibited a similar distribution pattern to that of the original dataset (Figure 8c) and in the updated dataset after SMOTE-ENN implementation (see Figure 8(d)). In general, the purpose of the HDPM is to minimize the errors during learning; thus, we expect that the HDPM performance can be enhanced from the balanced training datasets.

D. XGBOOST-BASED MACHINE LEARNING ALGORITHM (MLA) AND EVALUATION METRICS

After we balanced the training datasets, the MLA is used to learn and generate the HDPM. We used the extreme gradient boosting (XGBoost) algorithm to detect the presence or absence of heart disease. XGBoost is a type of supervised machine learning used for classification and regression modelling [45]. XGBoost is an enhanced algorithm based on the implementation of gradient boosting DTs with several modifications in terms of regularization, loss function and column sampling. Gradient boosting is a technique in which new models are created and used to predict the error or residuals, after which the scores are summed to get the final prediction result. The gradient descent method is used to minimize the loss score when new models are created. The objective function needs to be used to measure the model performance, which consists of two parts: training loss and regularization. The regularization term penalizes the complexity of the model and prevents overfitting. The objective function (loss function and regularization) can be presented as follows.

$$L(\phi) = \sum_i l\left(\hat{y}_{i,}, y_i\right) + \sum_k \Omega\left(f_k\right);$$

where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \quad (1)$$

The term l here is the differentiable convex loss function that calculates the difference between the prediction $\hat{y}_i$ and the target y_i. While the regularized term Ω penalizes the complexity of the model and the number of leaves in the tree are represented using T. Furthermore, each f_k corresponds to an independent tree structure q and leaf weight w. Finally, the term γ corresponds to the threshold and pre-pruning is performed while optimizing to limit the growth of the tree and λ is used to smooth the final learned weights to prevent overfitting.

We implemented XGBoost using the XGBoost V0.81 python library. The outlier data from heart disease training datasets are eliminated by using the DBSCAN method, and SMOTE-ENN is used to balance the training dataset. Finally, XGBoost is used to learn from the training dataset and generate the HDPM. We measured five performance metrics to compare the performance of the proposed model with that of state-of-the-art models and previous study results. In addition, we ensured the applicability of the proposed model by implementing the model into the HDCDSS to diagnose the subjects based on their current condition.

We used five performance metrics to evaluate the performance of the proposed model. A confusion matrix was used to measure four different potential outputs from the model: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP and TN outputs are defined as the number of subjects correctly classified as ''positive'' (presence of heart disease) and ''negative'' (healthy/ absence of heart disease), respectively, and FP and FN outputs as the number of subjects incorrectly classified as ''positive'' (presence of heart disease) when they are actually ''negative'' (healthy/ absence of heart disease) and incorrectly classified as ''negative'' (healthy/ absence of heart disease) when they are actually ''positive'' (presence of heart disease), respectively. We employed 10-fold cross validation to generate the models for all classification models, with

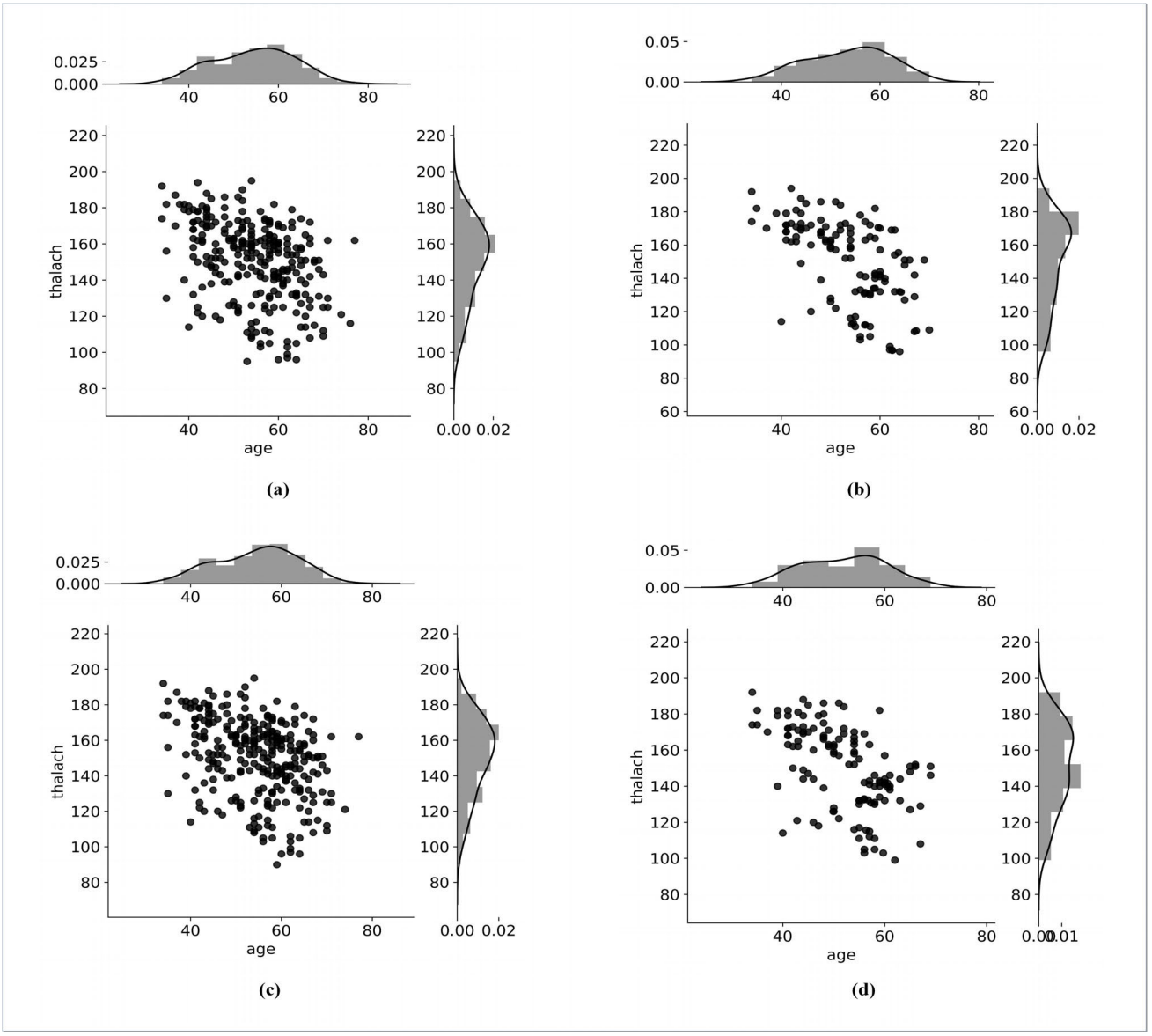

FIGURE 8. Data distribution of attributes "*age*" and "*thalach*" before and after SMOTE-ENN implementation for each dataset I (Statlog) (a), (b) and dataset II (Cleveland) (c), (d), respectively.

the final performance metric being the average. We implemented all the classification models in Python V3.6.5 by utilizing three libraries: sklearn V0.20.2, imbalanced-learn V0.4.3 and XGBoost V0.81. We performed the experiments on a computer with Intel Core i7-4790 (3.60 GHz × 8 cores), 16 GB RAM that runs with Windows 10 Pro 64-bit. The sklearn library is an open source python programming tool for machine learning, the imbalanced-learn library is also an open source python tool-box that consists of several methods to deal with imbalanced data, and the XGBoost library is an open source tool that implements the XGBoost algorithms in several programming languages, including Python. To simplify the implementation of the experimentations, we used default parameters provided by sklearn, imbalanced-learn and XGBoost. In addition, the following five performance metrics are measured. Accuracy (*acc*) is calculated as

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

precision (*pre*) is calculated as

$$pre = \frac{TP}{TP + FP}, \tag{3}$$

recall/sensitivity/true positive rate (*rec/sen/TPR*) is calculated as

$$rec = \frac{TP}{TP + FN}, \tag{4}$$

TABLE 5. Performance evaluation for dataset I (Statlog).

Model	Performance evaluation							
	acc (%)	*pre* (%)	*rec/sen/TPR* (%)	*f* (%)	*MCC*	*FPR (%)*	*FNR (%)*	*TNR (%)*
NB	84.07 ± 4.70	84.36 ± 7.85	80.00 ± 9.28	81.61 ± 5.46	0.68 ± 0.10	12.67 ± 7.57	20.00 ± 9.28	87.33 ± 7.57
LR	84.81 ± 4.21	85.49 ± 7.38	80.83 ± 11.81	82.21 ± 5.71	0.70 ± 0.08	12.00 ± 7.18	19.17 ± 11.81	88.00 ± 7.18
MLP	85.56 ± 4.21	86.12 ± 6.52	81.67 ± 12.25	82.99 ± 6.03	0.67 ± 0.12	11.33 ± 6.00	18.33 ± 12.25	88.67 ± 6.00
SVM	69.63 ± 7.37	72.90 ± 11.29	50.83 ± 10.83	59.52 ± 10.17	0.38 ± 0.16	15.33 ± 7.33	49.17 ± 10.83	84.67 ± 7.33
DT	74.81 ± 8.57	74.28 ± 13.61	70.83 ± 12.50	71.39 ± 9.27	0.49 ± 0.17	3.33 ± 12.74	28.33 ± 11.90	76.67 ± 12.74
RF	82.96 ± 8.15	85.15 ± 10.71	75.83 ± 12.05	79.64 ± 9.70	0.68 ± 0.14	12.00 ± 8.33	23.33 ± 11.06	88.00 ± 8.33
Proposed HDPM	**95.90 ± 5.55**	**97.14 ± 5.71**	**94.67 ± 11.08**	**95.35 ± 6.52**	**0.92 ± 0.10**	4.52 ± 6.94	3.33 ± 6.67	95.48 ± 6.94

Note: *acc* = accuracy, *pre* = Precision, *rec/sen/TPR* = recall/ sensitivity/ true positive rate, *f* = f-measure, *MCC* = Matthews correlation coefficient, FPR = false positive rate, FNR = false negative rate, TNR = true negative rate.

F-measure (f) is calculated as

$$f = \frac{2pr}{p+r}, \tag{5}$$

and MCC is calculated as

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

The value of MCC ranges from -1 to $+1$, which represent the performance of the classification model. The best model is achieved when the value of MCC is close or equal to $+1$ while the worst model is close or equal to -1. In addition, we also used the value of the area under the receiver operating characteristic curve (AUC) to compare the performance of the proposed model with that of other existing models. For the given k training data, the AUC can be calculated as [46], [47]

$$AUC\left(x^+, x^-\right) = \frac{1}{k^+k^-}\sum_{i=1}^{k^+}\sum_{j=1}^{k^-} 1_{h\left(x_i^+\right)>h\left(x_j^-\right)}, \tag{7}$$

where the term $1_{h\left(x_i^+\right)>h\left(x_j^-\right)}$ corresponds to a '1' when the elements $h\left(x_i^+\right) > h\left(x_j^-\right), \forall i = 1, 2, \ldots, ^+, \forall j = 1, 2, \ldots, k^-$, and '0' otherwise. The best model is achieved when the value of AUC is close or equal to 1. Additionally, we presented several additional metrics to measure the performance of the model such as false positive rate (*FPR)*, false negative rate (FNR), and true negative rate (*TNR*). FPR is used to represent the false alarm which the positive prediction result (presence of heart disease) will be given when the actual prediction output value is negative (absence of heart disease). The FPR can be calculated as

$$FPR = \frac{FP}{FP+TN}. \tag{8}$$

We used the FNR to represent the miss rate which is the probability that a positive prediction result will be missed by the test. The FNR can be calculated as

$$FNR = \frac{FN}{FN+TP}. \tag{9}$$

Finally, the TNR or specificity is used to show the probability that the actual negative subjects will test negative. The TNR can be calculated as

$$TNR = \frac{TN}{TN+FP}. \tag{10}$$

IV. RESULTS AND DISCUSSIONS

A. PERFORMANCE EVALUATION OF PROPOSED HDPM

The proposed HDPM was applied to both datasets and showed positive results for increasing the prediction accuracy as compared to other models. We selected six state-of-the-art MLAs (NB, LR, MLP, SVM, DT, and RF) that have been widely used in the research community and have a proven track record for accuracy and efficiency for comparison. We performed 10-fold cross-validation for all models and collected eight performance metrics: accuracy (*acc*), precision (*pre*), recall/sensitivity/true positive rate (*rec/sec/TPR*), f-measure (f), *MCC,* false positive rate (*FPR)*, false negative rate (*FNR*), and true negative rate (*TNR*). The findings revealed that the proposed model outperformed other models by achieving *acc, pre, rec/sec, f* up to 95.90%, 97.14%, 94.67%, 95.35% for dataset I and 98.40%, 98.57%, 98.33%, 98.32% for dataset II, respectively. In term of *MCC*, the proposed HDPM achieved the highest *MCC* value up to 0.92 and 0.97 for datasets I and II, respectively, which confirms the superiority of our proposed model relative to other models. In addition, in terms of false positive rate (*FPR*) and true positive rate (*TNR*), the results revealed that the proposed model achieved lowest *FPR* and highest *TNR* as compared with other models. The proposed model achieved *FPR* and high *TNR* by up to 4.52%, 95.48% and 1.67%, 98.33% for dataset I and II, respectively. The low *FPR* and high *TNR* value of the proposed model represented the capability of the HDPM model to minimize miss-rate and optimize prediction accuracy for both negative and positive subjects. The detailed performance results are presented in Table 5 and 6 for datasets I and II, respectively.

We further investigated the performance of the proposed HDPM using a receiver operating characteristic (ROC) curve visualization since a previous study [48] has used it to evaluate and illustrate the diagnostic capability as its threshold is changed. The ROC curve consists of the TP rate as the y-axis and FP rate as the x-axis with the area under the ROC curve (AUC) being calculated to show the performance of the

TABLE 6. Performance evaluation for dataset II (Cleveland).

Model	Performance evaluation							
	acc (%)	*pre* (%)	*rec/sen/TPR* (%)	*f* (%)	*MCC*	*FPR* (%)	*FNR* (%)	*TNR* (%)
NB	83.17 ± 7.64	84.18 ± 9.75	78.79 ± 8.29	81.25 ± 8.29	0.66 ± 0.15	13.12 ± 8.59	21.21 ± 8.29	86.88 ± 8.59
LR	84.85 ± 6.91	86.12 ± 7.85	80.22 ± 9.30	82.90 ± 7.80	0.70 ± 0.14	11.25 ± 6.73	19.78 ± 9.30	88.75 ± 6.73
MLP	84.15 ± 7.76	85.01 ± 9.74	80.22 ± 9.83	82.28 ± 8.66	0.68 ± 0.12	14.37 ± 9.29	18.41 ± 9.22	85.62 ± 9.29
SVM	71.06 ± 6.16	74.65 ± 9.51	59.23 ± 14.88	64.53 ± 9.43	0.43 ± 0.12	18.75 ± 10.46	40.77 ± 14.88	81.25 ± 10.46
DT	76.09 ± 4.86	74.21 ± 7.29	75.16 ± 8.00	74.31 ± 5.18	0.52 ± 0.09	24.38 ± 8.12	27.80 ± 7.31	75.62 ± 8.12
RF	82.14 ± 6.84	83.69 ± 8.63	76.54 ± 10.09	79.63 ± 8.36	0.66 ± 0.13	12.50 ± 8.39	22.03 ± 10.27	87.50 ± 8.39
Proposed HDPM	**98.40 ± 3.21**	**98.57 ± 4.29**	**98.33 ± 5.00**	**98.32 ± 3.37**	**0.97 ± 0.06**	1.67 ± 5.00	0.00 ± 0.00	98.33 ± 5.00

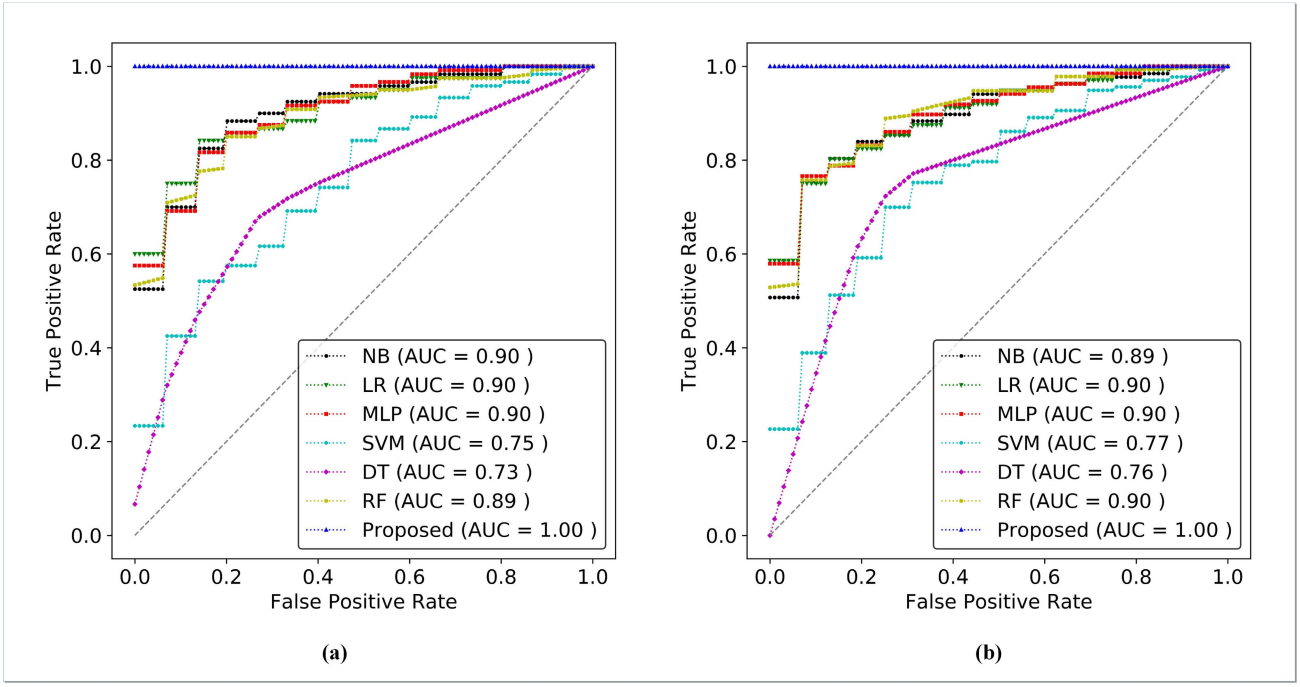

FIGURE 9. ROC curve visualization to compare the proposed model with other models for datasets (a) I (Statlog) and (b) II (Cleveland).

model. The best model is achieved when the value of AUC is close or equal to 1. Figure 9 shows that the proposed HDPM achieved higher AUC score than that of other models of up to 1.00 and 1.00 for datasets I and II, respectively, which confirmed that the proposed model outperformed other state-of-the-art models.

In addition, we followed a previous study [49] to evaluate the performance of the model using statistical-based significance testing to prove the significance of our proposed HDPM as compared with other state-of-the-art models. The paired t-test [50], [51] was applied to statistically test the significance between the proposed HDPM and other state-of-the-art models. We defined $h = 0$, i.e., the null hypothesis, as being no significance different between the proposed HDPM and other existing models. We performed 10-fold cross validation to collect ten accuracy data for all the models in Python V3.6.5 and applied the paired t-test using Scipy V1.2.0 library. We defined the significance level $= 0.05$, t (tabulated) $= 2.78$ and collected the h, p-value, and t (calculated) values for all datasets. The null hypothesis is accepted when the paired t-test return value of $h = 0$, and the null hypothesis is rejected if $h = 1$, which indicates a significance different between the proposed HDPM and the existing one. This could be supported by evidence that the p-value is less than the significance level (0.05) and t (calculated) is greater than t (tabulated). In Table 7, showing the paired t-test result for both datasets, the proposed HDPM is significantly different from the other models since for all datasets, $h = 1$, p-value $<$ significance level, and t (calculated) $> t$ (tabulated). Therefore, the proposed model has significant different as compared with other state-of-the-art models.

B. BENCHMARK WITH PREVIOUS STUDY RESULTS

In this section, we performed comparison study of our proposed HDPM with the results from previous studies. It should

TABLE 7. The results of paired t-test for datasets I (Statlog) and II (Cleveland).

Paired t-test		Dataset I (Statlog)	Dataset II (Cleveland)
Proposed HDPM vs. NB	*h*	1	1
	p-value	0.0001	0.0
	t (calculated)	4.88	5.516
Proposed HDPM vs. LR	*h*	1	1
	p-value	0.0002	0.0
	t (calculated)	4.776	5.333
Proposed HDPM vs. MLP	*h*	1	1
	p-value	0.0003	0.0001
	t (calculated)	4.457	5.091
Proposed HDPM vs. SVM	*h*	1	1
	p-value	0.0	0.0
	t (calculated)	8.546	11.816
Proposed HDPM vs. DT	*h*	1	1
	p-value	0.0	0.0
	t (calculated)	6.194	11.486
Proposed HDPM vs. RF	*h*	1	1
	p-value	0.001	0.0
	t (calculated)	3.937	6.461

TABLE 8. Benchmark with previous study results for dataset I (Statlog).

Author	Technique	Performance evaluation					
		acc (%)	*pre* (%)	*rec/sen* (%)	*f* (%)	*MCC*	*AUC*
Long et al. (2015) [13]	CFARS-AR	88.3	-	84.9	-	-	-
Nahato et al. (2015) [14]	RS-BPNN	90.40	-	94.67	-	-	0.92
Dwivedi (2018) [31]	LR	85	85	89	87	-	-
Amin et al. (2019) [32]	Vote with NB and LR	87.41	-	-	-	-	-
Proposed HDPM	**DBSCAN + SMOTE-ENN + XGBOOST**	**95.90**	**97.14**	**94.67**	**95.35**	**0.92**	**1.00**

TABLE 9. Benchmark with previous study results for dataset II (Cleveland).

Author	Technique	Performance evaluation					
		acc (%)	*pre* (%)	*rec/sen* (%)	*f* (%)	*MCC*	*AUC*
Verma et al. (2016) [15]	CFS + PSO + K-means + MLP	90.28	-	-	-	-	-
Haq et al. (2018) [16]	Relief + LR	89	-	77	-	0.89	0.88
Saqlain et al. (2019) [17]	MFSFSA + SVM	81.19	-	72.92	-	0.85	-
Latha and Jeeva (2019) [19]	Majority vote with NB, BN, RF and MP	85.48	-	-	-	-	-
Ali et al. (2019) [20]	Stacked SVMs	92.22	-	82.92	-	0.85	-
Mohan et al. (2019) [21]	HRFLM	88.4	90.1	92.8	90	-	-
Gupta et al. (2020) [18]	FAMD + RF	93.44	-	89.28	92.59	0.87	0.93
Proposed HDPM	**DBSCAN + SMOTE-ENN + XGBOOST**	**98.40**	**98.57**	**98.33**	**98.32**	**0.97**	**1.00**

be noted that since we utilized the same datasets, we directly took the results from previous studies without implementing their techniques. The detailed comparison results with previous studies for datasets I and II are given in Table 8 and 9, respectively.

Previous studies have utilized the Statlog dataset for generating the machine learning model to diagnose the heart disease. Long *et al.* (2015) [13] proposed the CFARS-AR and achieved $acc = 88.3\%$ and $rec/sen = 84.9\%$. Nahato *et al.* (2015) [14] used the rough set method with RS-BPNN and achieved $acc = 90.40\%$, $rec/sen = 94.67\%$ and $AUC = 0.92$. Dwivedi (2018) [31] used LR and achieved $acc = 85\%$, $pre = 85\%$, $rec/sec = 89\%$, and $f = 87\%$. Amin *et al.* (2019) [32] utilized the voting method with NB and LR and achieved $acc = 87.41\%$. The proposed HDPM achieved $acc = 95.90\%$, $pre = 97.14\%$, $rec/sec = 94.67\%$, $f = 95.35\%$, $MCC = 0.92$, and $AUC = 1.00$. In terms of accuracy, the proposed HDPM achieved the highest accuracy with an average improvement of 8.12% as compared with previous study results. Overall, we can conclude that our proposed method outperformed all the previous study results in terms of accuracy, f-measure, MCC and AUC.

In addition, several researchers have also used the Cleveland dataset to predict heart disease. Verma *et al.* (2016) [15] developed a hybrid model with CFS selection, PSO, K-means clustering and MLP and achieved $acc = 90.28\%$. Haq *et al.* (2018) [16] proposed a hybrid system using Relief-based feature selection and LR, and achieved $acc = 89\%$, $rec/sec = 77\%$, $MCC = 0.89$, and $AUC = 0.88$. Saqlain *et al.* (2019) [17] used MFSFSA and SVM and achieved $acc = 81.19\%$, $rec/sec = 72.92\%$, and $MCC = 0.85$. Latha and Jeeva (2019) [19] used majority voting with NB, BN, RF, and MLP and achieved $acc = 85.48\%$. Ali *et al.* (2019) [20]

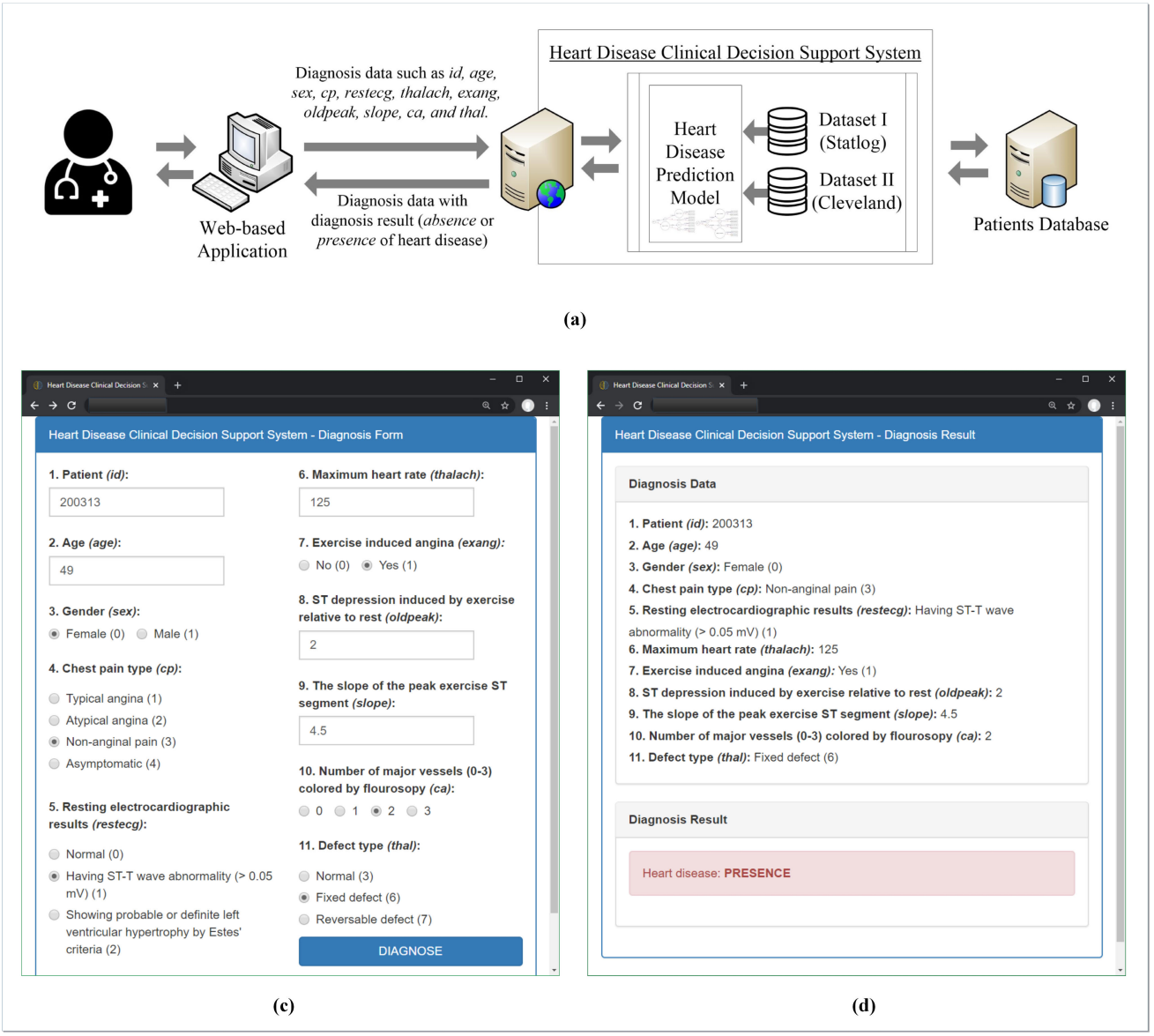

FIGURE 10. Heart Disease Clinical Decision Support System (HDCDSS) (a) architecture framework, (b) diagnosis form, and (c) diagnosis result.

used stacked SVMs and achieved $acc = 92.22\%$, $rec/sec = 82.92\%$, and $MCC = 0.85$. Mohan *et al.* (2019) [21] developed HRFLM and achieved $acc = 88.4\%$, $pre = 90.1\%$, $rec/sec = 92.8\%$, and $f = 90\%$. Gupta *et al.* (2020) [18] utilized the FAMD-based feature extraction and RF algorithm and achieved $acc = 93.44\%$, $rec/sec = 89.28\%$, $f = 92.59\%$, $MCC = 0.87$, and $AUC = 0.93$. Finally, the proposed HDPM achieved $acc = 98.40\%$, $pre = 98.57\%$, $rec/sec = 98.33\%$, $f = 98.32\%$, $MCC = 0.97$, and $AUC = 1.00$. In terms of accuracy, the proposed HDPM achieved the highest accuracy with an average improvement of 9.83% as compared with previous study results. Overall, we can conclude that our proposed method outperformed all the previous study results in all six-performance metrics (*acc*, *pre*, *rec/sen*, *f* , *MCC*, and *AUC*).

It should be noted that a direct comparison of the presented results is not fair since they have been derived by different data pre-processing and training/testing approaches. In addition, the prediction model performance depends on several factors such as features selections, data types and its size, noise filtering, hyperparameters, data sampling, model selection, etc. Therefore, these general comparison (as presented in Table 8 and 9) cannot be used as the main evidence to conclude the performance of given prediction models but it can be used simply as a general comparison between the proposed HDPM and previous studies.

V. APPLICATION FOR THE HEART DISEASE CLINICAL DECISION SUPPORT SYSTEM (HDCDSS)

The prototype of the web-based Heart Disease Clinical Decision Support System (HDCDSS) was developed to provide a simple and convenient way for medical clinicians to diagnose subjects/patients based on their current condition.

The HDCDSS was developed in Python V3.6.5 by utilizing Flask V1.0.2 as a Python Web Server Gateway Interface (WSGI) with Bootstrap V3.3.7 for data representation, while the proposed HDPM was loaded using Joblib V0.14.1 and XGBoost V0.81. The patients' data and the prediction results were stored into MongoDB by using Pymongo V3.7.1. MongoDB was selected since it has been widely adopted in the healthcare field [52], [53]. As illustrated in Figure 10(a), clinicians can access the HDCDSS through their web-browser in the same local network since the medical data are confidential information and cannot be stored in the cloud. The personal data such as patient id (*id*), *age*, and *gender* are then combined with the diagnosis data, such as resting electrocardiographic result (*restecg*), maximum heart rate (*thalach*), exercise induced angine (*exang*), ST depression induced by exercise relative to rest (*oldpeak*), slope of the peak exercise ST segment (*slope*), number of major vessels (0-3) colored by fluoroscopy (*ca*), and defect type (*thal*), and then transmitted into a secure web server through an application programming interface (API) and stored in a database. The proposed HDPM generated from datasets I (Statlog) and II (Cleveland) is then used to predict the subjects' heart disease status based on the inputted data, and the prediction result is then sent back to the HDCDSS's diagnosis result interface.

Figure 10(b) shows the HDCDSS diagnosis form in which clinicians can fill out the patients' information, including their currents conditions. Once all the input fields are filled, the user can press the ''diagnose'' button to send all the data to the secure web server, which loads the trained proposed HDPM to diagnose the subjects' heart disease status. Figure 10(c) shows the diagnosis result interface after sending the data to the web server. The result includes the previously submitted data and the status (presence or absence) of heart disease. The developed HDCDSS is expected to help clinicians to diagnose patients and improving heart disease clinical decision making effectively and efficiently. Therefore, early treatment could be conducted to prevent the deaths caused by late heart disease diagnosis. This prototype/demonstration is only limited to the specific datasets; therefore, the trained prediction model cannot be applied for other demographic patients/subjects. Once we have collected more complex datasets, it could improve the predictive performance for wider demographic patients/subjects. In addition, we have not applied the developed model in the clinical trial due to limitation of the dataset. In our case, we have used the dataset based on specific demographic patient (USA). The clinical trial could be applied to our model once we gather another demographic patient (for example in Korea) and it is beyond the scope of our current study.

VI. CONCLUSION

We proposed an effective heart disease prediction model (HDPM) for heart disease diagnosis by integrating DBSCAN, SMOTE-ENN, and XGBoost-based MLA to improve prediction accuracy. The DBSCAN was applied to detect and remove the outlier data, SMOTE-ENN was used to balance the unbalanced training dataset and XGBoost MLA was adopted to learn and generate the prediction model. Two publicly available datasets of heart disease were utilized by produce the generalized prediction model. We performed evaluation analysis of our proposed model with other classification models and the results from previous studies. In addition, we presented the statistical evaluation to confirm the significant of our model as compared to other models. The experimental results confirmed that the proposed model achieved better performance than that of state-of-the-art models and previous study results, by achieving an accuracy up to 95.90% and 98.40% for datasets I and II, respectively. In addition, the statistical-based analysis result also showed the significant improvement for the proposed model as compared with the other models.

Furthermore, we also designed and developed the proposed HDPM into the Heart Disease Clinical Decision Support System (HDCDSS) to diagnose the subjects'/patients' heart disease status effectively and efficiently. The HDCDSS gathered the patient data combined with other diagnosis data and transmitted them to a secure web server. All the transmitted diagnosis data were then stored into MongoDB, which can effectively provide timely response with rapidly increasing medical data. The proposed HDPM was then loaded to diagnose the patients' current heart disease status, which was later sent back to the HDCDSS's diagnosis result interface. Thus, the developed HDCDSS is expected to help clinicians to diagnose patients and improving heart disease clinical decision making effectively and efficiently. Finally, the overall designed and developed HDCDSS in this study can be used as a practical guideline for the healthcare practitioners.

In the future, we will consider the comparison of other data sampling with the model hyper-parameters and broader medical datasets. In addition, a comparison and analysis study with different outlier detection methods could be further investigated. Furthermore, with the increasing concerns about privacy, security and time-sensitive applications, edge computing and edge device concepts could be further studied with the goal of improving the medical clinical decision support system. In this study, we have not obtained any feedback from heart specialist yet. In the future, once specific demographic dataset (from Korea) is collected, the comments from local heart specialist for verifying dataset and prediction model could be presented.

ACKNOWLEDGMENT

This article is a tribute made of deep respect of a wonderful person, friend, advisor, and supervisor, Yong-Han Lee (1965–2017).

REFERENCES

[1] World Health Organization. (2017). *Cardiovascular Diseases (CVDs)*. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases/

[2] E. J. Benjamin *et al.*, ''Heart disease and stroke statistics—2019 update: A report from the American heart association,'' *Circulation*, vol. 139, no. 10, pp. e56–e528, Mar. 2019, doi: 10.1161/CIR.0000000000000659.

[3] Statistics Korea. (2018). *Causes of Death Statistics in 2018*. [Online]. Available: http://kostat.go.kr/portal/eng/pressReleases/8/10/index.board?bmode=read&bSeq=&aSeq=378787

[4] World Health Organization. (2017). *Cardiovascular Diseases (CVDs)*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[5] P. Greenland, J. S. Alpert, G. A. Beller, E. J. Benjamin, M. J. Budoff, Z. A. Fayad, E. Foster, M. A. Hlatky, J. M. Hodgson, F. G. Kushner, M. S. Lauer, L. J. Shaw, S. C. Smith, A. J. Taylor, W. S. Weintraub, and N. K. Wenger, "2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: A report of the American college of cardiology foundation/American heart association task force on practice guidelines," *Circulation*, vol. 122, no. 25, pp. e584–e636, Dec. 2010, doi: 10.1161/CIR.0b013e3182051b4c.

[6] J. Perk *et al.*, "European guidelines on cardiovascular disease prevention in clinical practice (version 2012): The fifth joint task force of the European society of cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of nine societies and by invited experts) ∗ developed with the special contribution of the European association for cardiovascular prevention & rehabilitation (EACPR)," *Eur. Heart J.*, vol. 33, no. 13, pp. 1635–1701, Jul. 2012, doi: 10.1093/eurheartj/ehs092.

[7] G.-M. Park and Y.-H. Kim, "Model for predicting cardiovascular disease: Insights from a Korean cardiovascular risk model," *Pulse*, vol. 3, no. 2, pp. 153–157, 2015, doi: 10.1159/000438683.

[8] G. J. Njie, K. K. Proia, A. B. Thota, R. K. C. Finnie, D. P. Hopkins, S. M. Banks, D. B. Callahan, N. P. Pronk, K. J. Rask, D. T. Lackland, and T. E. Kottke, "Clinical decision support systems and prevention," *Amer. J. Preventive Med.*, vol. 49, no. 5, pp. 784–795, Nov. 2015, doi: 10.1016/j.amepre.2015.04.006.

[9] V. Sintchenko, E. Coiera, J. R. Iredell, and G. L. Gilbert, "Comparative impact of guidelines, clinical data, and decision support on prescribing decisions: An interactive Web experiment with simulated cases," *J. Amer. Med. Inform. Assoc.*, vol. 11, no. 1, pp. 71–77, Jan. 2004, doi: 10.1197/jamia.M1166.

[10] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, "Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success," *BMJ*, vol. 330, no. 7494, p. 765, Apr. 2005, doi: 10.1136/bmj.38398.500764.8F.

[11] H. B. Bosworth, M. K. Olsen, T. Dudley, M. Orr, M. K. Goldstein, S. K. Datta, F. McCant, P. Gentry, D. L. Simel, and E. Z. Oddone, "Patient education and provider decision support to control blood pressure in primary care: A cluster randomized trial," *Amer. Heart J.*, vol. 157, no. 3, pp. 450–456, Mar. 2009, doi: 10.1016/j.ahj.2008.11.003.

[12] J. Hunt, J. Siemienczuk, W. Gillanders, B. LeBlanc, Y. Rozenfeld, K. Bonin, and G. Pape, "The impact of a physician-directed health information technology system on diabetes outcomes in primary care: A pre- and post-implementation study," *J. Innov. Health Inform.*, vol. 17, no. 3, pp. 165–174, Sep. 2009, doi: 10.14236/jhi.v17i3.731.

[13] N. C. Long, P. Meesad, and H. Unger, "A highly accurate firefly based algorithm for heart disease prediction," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 8221–8231, Nov. 2015, doi: 10.1016/j.eswa.2015.06.024.

[14] K. B. Nahato, K. N. Harichandran, and K. Arputharaj, "Knowledge mining from clinical datasets using rough sets and backpropagation neural network," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–13, Mar. 2015, doi: 10.1155/2015/460189.

[15] L. Verma, S. Srivastava, and P. C. Negi, "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data," *J. Med. Syst.*, vol. 40, no. 7, p. 178, Jul. 2016, doi: 10.1007/s10916-016-0536-z.

[16] 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, pp. 1–21, Dec. 2018, doi: 10.1155/2018/3860146.

[17] 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, doi: 10.1007/s10115-018-1185-y.

[18] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "MIFH: A machine intelligence framework for heart disease diagnosis," *IEEE Access*, vol. 8, pp. 14659–14674, 2020, doi: 10.1109/ACCESS.2019.2962755.

[19] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Inform. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100203, doi: 10.1016/j.imu.2019.100203.

[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, doi: 10.1109/ACCESS.2019.2909969.

[21] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[22] X. Liu, Q. Yang, and L. He, "A novel DBSCAN with entropy and probability for mixed data," *Cluster Comput.*, vol. 20, no. 2, pp. 1313–1323, Jun. 2017, doi: 10.1007/s10586-017-0818-3.

[23] C.-H. Lin, K.-C. Hsu, K. R. Johnson, M. Luby, and Y. C. Fann, "Applying density-based outlier identifications using multiple datasets for validation of stroke clinical outcomes," *Int. J. Med. Inform.*, vol. 132, Dec. 2019, Art. no. 103988, doi: 10.1016/j.ijmedinf.2019.103988.

[24] Z. H. Ismail, A. K. K. Chun, and M. I. S. Razak, "Efficient herd—Outlier detection in livestock monitoring system based on density—Based spatial clustering," *IEEE Access*, vol. 7, pp. 175062–175070, 2019, doi: 10.1109/ACCESS.2019.2952912.

[25] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[26] T. Le, M. Lee, J. Park, and S. Baik, "Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset," *Symmetry*, vol. 10, no. 4, p. 79, Mar. 2018, doi: 10.3390/sym10040079.

[27] T. Le and S. Baik, "A robust framework for self-care problem identification for children with disability," *Symmetry*, vol. 11, no. 1, p. 89, Jan. 2019, doi: 10.3390/sym11010089.

[28] T. Le, M. T. Vo, B. Vo, M. Y. Lee, and S. W. Baik, "A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction," *Complexity*, vol. 2019, pp. 1–12, Aug. 2019, doi: 10.1155/2019/8460934.

[29] *Statlog (Heart) Data Set*. Accessed: Oct. 2, 2019. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/statlog+(heart)

[30] *Heart Disease Data Set*. Accessed: Oct. 2, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[31] 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, May 2018, doi: 10.1007/s00521-016-2604-1.

[32] 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, doi: 10.1016/j.tele.2018.11.007.

[33] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artif. Intell. (IJCAI)*, Montreal, QC, Canada, vol. 2, Aug. 1995, pp. 1137–1145. [Online]. Available: http://ijcai.org/Proceedings/95-2/Papers/016.pdf

[34] R. Kohavi and D. Wolpert, "Bias plus variance decomposition for zero-one loss functions," in *Proc. 13th Int. Conf. Mach. Learn.*, San Francisco, CA, USA, 1996, pp. 275–283.

[35] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Diego, CA, USA: Elsevier, 2012.

[36] *Weka 3: Data Mining Software in Java*. Accessed: Oct. 19, 2019. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/

[37] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, pp. 245–271, Dec. 1997, doi: 10.1016/S0004-3702(97)00063-5.

[38] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[39] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, Portland, OR, USA, 1996, pp. 226–231.

[40] M. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, "Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest," *Appl. Sci.*, vol. 8, no. 8, p. 1325, Aug. 2018, doi: 10.3390/app8081325.

[41] G. Alfian, M. Syafrudin, and J. Rhee, "Real-time monitoring system using smartphone-based sensors and NoSQL database for perishable supply chain," *Sustainability*, vol. 9, no. 11, p. 2073, Nov. 2017, doi: 10.3390/su9112073.

[42] M. Syafrudin, G. Alfian, N. Fitriyani, and J. Rhee, "Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing," *Sensors*, vol. 18, no. 9, p. 2946, Sep. 2018, doi: 10.3390/s18092946.
[43] M. Syafrudin, N. Fitriyani, G. Alfian, and J. Rhee, "An affordable fast early warning system for edge computing in assembly line," *Appl. Sci.*, vol. 9, no. 1, p. 84, Dec. 2018, doi: 10.3390/app9010084.
[44] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 559–563, Jan. 2017.
[45] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[46] C. Marrocco, R. P. W. Duin, and F. Tortorella, "Maximizing the area under the ROC curve by pairwise feature combination," *Pattern Recognit.*, vol. 41, no. 6, pp. 1961–1974, Jun. 2008, doi: 10.1016/j.patcog.2007.11.017.
[47] K.-A. Toh, J. Kim, and S. Lee, "Maximizing area under ROC curve for biometric scores fusion," *Pattern Recognit.*, vol. 41, no. 11, pp. 3373–3392, Nov. 2008, doi: 10.1016/j.patcog.2008.04.002.
[48] S. H. Jee *et al.*, "A coronary heart disease prediction model: The korean heart study," *BMJ Open*, vol. 4, no. 5, May 2014, Art. no. e005025, doi: 10.1136/bmjopen-2014-005025.
[49] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, D. S. Rajput, R. Kaluri, and G. Srivastava, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," *Evol. Intell.*, vol. 13, no. 2, pp. 185–196, Nov. 2019, doi: 10.1007/s12065-019-00327-1.
[50] B. R. Kirkwood, J. A. C. Sterne, and B. R. Kirkwood, *Essential Medical Statistics*, 2nd ed. Malden, MA, USA: Blackwell Science, 2003.
[51] M. Xu, D. Fralick, J. Z. Zheng, B. Wang, X. M. Tu, and C. Feng, "The differences and similarities between two-sample T-test and paired T-test," *Shanghai Arch, Psychiatry*, vol. 29, no. 3, pp. 184–188, Jun. 2017, doi: 10.11919/j.issn.1002-0829.217070.
[52] G. Alfian, M. Syafrudin, M. Ijaz, M. Syaekhoni, N. Fitriyani, and J. Rhee, "A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing," *Sensors*, vol. 18, no. 7, p. 2183, Jul. 2018, doi: 10.3390/s18072183.
[53] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Development of disease prediction model based on ensemble learning approach for diabetes and hypertension," *IEEE Access*, vol. 7, pp. 144777–144789, 2019, doi: 10.1109/ACCESS.2019.2945129.

NORMA LATIF FITRIYANI received the bachelor's degree from Universitas Islam Negeri Sunan Kalijaga, Yogyakarta, Indonesia, and the master's degree from the National Taiwan University of Science and Technology, Taipei, Taiwan. She is currently pursuing the Ph.D. degree with the Department of Industrial and Systems Engineering, Dongguk University, Seoul, South Korea. She has published numerous research articles in several international peer-reviewed journals, including IEEE Access, *Food Control*, *Sensors*, *Applied Sciences*, *Asia Pacific Journal of Marketing and Logistics*, and *Sustainability*. Her research interests include health informatics, machine learning, the Internet of Things, sensors, and image processing.

MUHAMMAD SYAFRUDIN received the bachelor's degree from Universitas Islam Negeri Sunan Kalijaga, Yogyakarta, Indonesia, and the Ph.D. degree from Dongguk University, Seoul, South Korea. He is currently an Assistant Professor with the Department of Industrial and Systems Engineering, Dongguk University. He is also an Instructor with popular practical course on undergraduate topics in programming languages and database systems. He has collaborated actively with researchers in several other disciplines of engineering, particularly information processing and machine learning on problems at the real-world industrial applications. He was selected and invited to participate with the World Class Scholar Symposium (SCKD) Event hosted by the Ministry of Research, Technology, and Higher Education, Indonesia, in August 2019, to make contributions on accelerating the Indonesian national development. He was recognized for Excellence in research, teaching, and outreach. He has published numerous research articles in several international peer-reviewed journals, including IEEE Access, *Food Control*, *Sensors*, *Applied Sciences*, *Asia Pacific Journal of Marketing and Logistics*, and *Sustainability*. His research interests include industrial artificial intelligence, machine learning, information systems, edge-computing, the Internet of Things, big data, health informatics, and smart factory ranging from theory to design and implementation. He serves as a Reviewer Board Members for *Sensors* and *Algorithms* (MDPI) and a Review Editor for the *IoT* and *Sensor Networks* (Frontiers in Communications and Networks).

GANJAR ALFIAN (Member, IEEE) received the B.Eng. degree from the Department of Informatics Engineering, Universitas Islam Negeri Sunan Kalijaga, Yogyakarta, Indonesia, in 2009, and the M.Eng. and Dr.Eng. degrees from the Department of Industrial and Systems Engineering, Dongguk University, Seoul, South Korea, in 2012 and 2016, respectively. He has been an Assistant Professor with the Industrial AI Research Center, Nano Information Technology Academy, Dongguk University, since 2016. In July 2017, he was a short-term Visiting Researcher with the VSB-Technical University of Ostrava, Czech Republic. He was recognized for Excellence in research, teaching, and outreach. He has published numerous research articles in several international peer-reviewed journals, including *Computers and Industrial Engineering*, *Journal of Food Engineering*, *Journal of Public Transportation*, IEEE Access, *Food Control*, *Sensors*, *Applied Sciences*, *Asia Pacific Journal of Marketing and Logistics*, and *Sustainability*. His research interests include machine learning, deep learning, RFID, the Internet of Things, big data, health informatics, and simulation and car sharing service. He was a recipient of the International Conference on Science and Technology (ICST) Best Paper Award, in 2019.

JONGTAE RHEE received the B.S. degree in industrial engineering from Seoul National University, the M.S. degree in industrial engineering from the Korea Advanced Institute of Science and Technology, and the Ph.D. degree in industrial engineering from the University of California, Berkeley. He is currently a Professor with the Department of Industrial and Systems Engineering and the Director of the Industrial Artificial Intelligence Research Center, Dongguk University, Seoul, South Korea. He has been leading researches and projects related to practical artificial neural network models, including for production and operation planning, personalized healthcare, and smart factory. He was recognized for Excellence in research, teaching, and outreach. He has published numerous research articles in several international peer-reviewed journals, including the IEEE Sensors Journal, *Expert Systems with Applications*, *Computers and Industrial Engineering*, *Journal of Food Engineering*, the *International Journal of Production Research*, *Journal of Public Transportation*, *Journal of Food Agriculture and Environment*, IEEE Access, *Food Control*, *Sensors*, *Applied Sciences*, *Asia Pacific Journal of Marketing and Logistics*, and *Sustainability*. His research interests include industrial artificial intelligence, machine learning, optimization, the Internet of Things, big data, and sensors.