Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 6913043, 14 pages
https://doi.org/10.1155/2022/6913043

Research Article

Anomaly Detection in Heart Disease Using a Density-Based Unsupervised Approach

Y. A. Nanehkaran,[1] **Zhu Licai**,[1] **Junde Chen**,[2] **Ahmed A. M. Jamel**,[3] **Zhao Shengnan**,[1] **Yahya Dorostkar Navaei**,[4] **and Mohsen Abdollahzadeh Aghbolagh**[5]

[1]*School of Information Engineering, Yancheng Teachers University, Yancheng, 224002 Jiangsu, China*
[2]*School of Informatics, Xiamen University, Xiamen, 361005 Fujian, China*
[3]*Netcom Bilgisayar A.S., Department of Research and Development, Melikgazi, Kayseri, Turkey*
[4]*Department of Computer and Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran*
[5]*School of Information Technology and Data Science, Irkutsk National Research University, Russia*

Correspondence should be addressed to Yahya Dorostkar Navaei; y.dorostkar@qiau.ac.ir

Received 6 November 2021; Accepted 3 March 2022; Published 26 March 2022

Academic Editor: B. B. Gupta

Cardiovascular disease is one of the most common diseases in the modern world, which, if diagnosed early, can greatly reduce the damage to the patient. Diagnosis of heart disease requires great care, and in some cases, the process can be disrupted by human error. Machine learning methods, especially data mining, have gained international acceptance in almost all aspects of life, especially the prediction of heart disease. On the other hand, datasets related to heart patients have many biological features that most of these features do not have a direct impact on diagnosis. By removing redundant features from the dataset, in addition to reducing computational complexity, the accuracy of heart patients' predictions can also be increased. This paper presents a density-based unsupervised approach to the diagnosis of abnormalities in heart patients. In this method, the basic features in the dataset are first selected based on the filter-based feature selection approach. Then, the DBSCAN clustering method with adaptive parameters has used to increase the clustering accuracy of healthy instances and to determine abnormal instances as cardiac patients. Partition clustering methods suffer from the selection of the number of clusters and the initial central points and are very sensitive to noise. The DBSCAN method solves these problems by creating density-based clusters, but the selection of the neighborhood radius threshold and the number of connected points in the neighborhood remains unresolved. In the proposed method, these two parameters are selected adaptively to achieve the highest accuracy for the diagnosis and prediction of heart patients. The results of the experiments show that the accuracy of the proposed method for predicting heart patients is approximately 95%, which has improved in comparison with previous methods.

1. Introduction

Cardiovascular disease is a disease that affects the heart or blood vessels. This is one of the most common diseases in the modern world. Heart disease accounts for twelve million deaths worldwide. In developed countries such as the United States, European countries, and Japan, the most prevalent cause of death in adults is cardiovascular disease [1, 2]. Cardiovascular diseases include a wide range of issues related to the human heart and its function. The diagnosis of heart disease must be made accurately and correctly. It is usually diagnosed by a medical professional. When timely prediction and studying patients' history and lifestyle have been considered thoroughly, cardiovascular diseases can be predicted and preventive measures can be taken to eliminate or suppress these life-threatening diseases [3, 4]. In case we use techniques integrated with the medical information system, this benefit will be greater and will cause cost reduction. Integrated systems try to diagnose the stages of the disease and prescribe the necessary reactions to prevent the progression of the disease by collecting useful information from the patient's history. Data mining has recently gained

international acceptance in almost all walks of life, and medicine is not an exception. Due to its advantages in extracting latent knowledge from raw data, data mining methods are suitable options for building integrated information systems to diagnose cardiovascular diseases. Due to the breadth of data mining in improving health care, various techniques have been proposed to diagnose and predict cardiovascular disease [5–8].

In the process of diagnosing heart disease, recognizing the basic features to determine the progress of heart disease is one of the important steps that can greatly improve the accuracy of the process of diagnosing and predicting heart disease. Abnormal specimens among heart patients refer to observations that have features significantly different from other specimens [9].

According to the significant and interesting insights that are often presented, anomaly detection technologies play an important role in various fields [10]. Considering the importance of basic features in patient label diagnosis, the discovery of anomalous specimens based on these features can be very effective in increasing the accuracy of disease diagnosis systems [11, 12]. This paper presents a density-based unsupervised approach to detect anomalies among heart patients.

The problems of partition clustering methods include selecting the number of clusters and determining the initial central point. The probability of error in clustering is high due to these problems. Therefore, in order to overcome the recent problems, the DBSCAN clustering method works based on the distance between the samples and the number of samples in the neighborhood radius. Thus, the DBSCAN clustering method does not need to determine the number of clusters and determine the initial point. Clusters are determined by the density of the data, and wherever there are a large number of similar samples, a cluster is formed. But there is still a problem with this method, determining the threshold for the minimum distance between samples and determining the number of samples in the neighborhood. In this paper, in order to solve this problem, the adaptive DBSCAN clustering approach has been used to identify and predict heart patients. In this method, these two parameters are selected in order to achieve high clustering accuracy, according to an optimization process. In other words, the proposed method adjusts these parameters adaptively to obtain near-optimal results for the diagnosis of heart patients.

This study uses the Heart Disease Prediction dataset [13] which is located in the UCI standard data repository. The proposed approach first extracts important features in the dataset that are more correlated with the class label of instructional samples. Attribute correlation with a class label refers to those features that can be used to determine the class label of instructional samples and those attributes with little effect on determining the instance class label. Thus, in addition to reducing the dimensions of the data, the prediction accuracy of the model also increases. In this research, after selecting the basic features, the density-based spatial clustering of applications with noise (DBSCAN) [14] with adaptive parameters has been applied in order to increase the clustering accuracy of normal samples and afterwards to determine anomalous samples. Based on the density of samples in the educational space and important features, this algorithm decides to connect samples close to each other and create clusters based on similar samples. The problems of prototype-based clustering methods, which include determining the number of clusters and deciding on the starting point for cluster centers, are solved in this method. At the same time, the parameters of this algorithm, which include the neighborhood radius and the minimum number of samples in the neighborhood to connect to each other, play a decisive role in the accuracy of clustering. Similar samples are connected based on fine-tuning of parameters and form clusters with different shapes. The number of clusters is also determined based on the density of data in different areas of the educational space. Finally, specimens that do not belong to any of the clusters are identified as anomalous specimens.

The main contribution of the present article is summarized as follows:

(1) Feature selection using a filter approach based on the correlation of features with the class label

(2) Adaptive adjustment of the neighborhood radius and number of data points in the neighborhood in density-based clustering

(3) Comparison with various classification methods

In the continuation of the article, in the second part, the previous works in the field of heart disease prediction will be reviewed. In the third section, the proposed method will be described in detail. In the fourth section, the test results and evaluation of the proposed method will be presented. Section 5 will provide conclusions and future work.

2. Review of the Related Literature

Medical Database is a large database that holds a variety of medical records such as treatment records, patient history, drug profiles, pathology reports, radiology reports, signals, and images. Medical care data are characterized by their complexity and diversity depending on the type of data. This data is large, indeterminate, and distributed and may be incorrect and unrelated, or it could have missing values. Detecting patterns in this type of data using traditional statistical methods is difficult if not unattainable. Therefore, data mining techniques in medical information have offered more effective analysis methods to improve diagnostic capabilities and patient care. Data mining in medical records is associated with the idea that there is more hidden knowledge in this data that is not readily available. To discover this knowledge, a set of techniques are commonly used in medical data, instead of a single method, for various purposes and to answer important medical questions, some of which will be examined below.

Singh and Rajesh used the extended K-means clustering method to cluster patients with heart disease. Clustering suffers from the problem of selecting the starting point in the algorithm results, in terms of both the number of clusters found and their centers. Methods for amplifying the K

-means clustering algorithm are discussed in this paper [14]. Liu et al. have proposed a hierarchical local density clustering algorithm based on the nearest inverse neighbors, RNN-LDH. Constructing and using the nearest inverse diagram, the extended core regions are found as the primary clusters. A new local density criterion has been defined to calculate the density of each sample [15]. Lima et al. have classified the heart rate images in various cardiovascular diseases. After using Info GAN architecture to produce artificial images related to anomalous classes, a two-dimensional permutation neural network has been proposed for classification [16].

Liu et al. have proposed a novel Single-Objective Generative Adversarial Active Learning (SO-GAAL) to identify anomalies, which can be used directly from potentially informative anomaly environments based on the min-max game between a generator and detector [10]. Talab et al. discuss a cost-effective and reliable method for diagnosing heart anomalies based on neural networks using mobile phones that are commonly available to every user today [17]. Umasankar and Thiagarasu provide a framework that includes the preprocessing phase, exploring the fuzzy associative law, and extracting the fuzzy correlation law for decision-making. In this paper, the proposed framework has focused mainly on the criteria that could possibly cause a heart attack among people [18]. Gokulnath and Shantharajah have proposed a support vector machine (SVM) optimization function in which the objective function in the genetic algorithm (GA) is used to select the most important features for heart disease [19]. Vivekanandan and Iyengar presented a performance analysis based on the developed and modified DE strategy. With selected important features, heart disease prediction is performed using fuzzy AHP and a leading neural network [20].

In Kavitha et al., a new machine learning approach based on a combination of a decision tree and a random forest is proposed to predict heart disease. In this study, the decision tree and random forest method alone have been implemented on the prediction data of heart patients [21]. Bharti et al. have implemented different machine learning algorithms and deep learning on a heart disease dataset to compare the results and analysis of them [22].

In El-Hasnony et al., five (MMC, random, adaptive, QUIRE, and AUDI) selection strategies for multilabel active learning were applied and used for reducing labelling costs by iteratively selecting the most relevant data to query their labels [23]. Qiu et al. focus on a new method of data augmentation to solve the data imbalance problem by using optimal transport. In this work, the ECG disease data from normal ECG beats to balance the data among different categories has augmented [24].

3. Methodology

As mentioned, the proposed method uses a density-based clustering approach based on parameter matching. In the proposed method, the dataset used includes several features that the use of all features to identify heart patients not only increases the computational complexity of the system but also reduces the accuracy of identifying and predicting heart patients. Therefore, in the first step of this method, the appropriate features will be selected based on the filter-based feature selection approach. In this approach, a threshold value is considered to select features or not. When attributes are selected based on this threshold, these attributes are sent to the training process as representative properties of the original dataset. In the proposed method, training on sick and healthy samples will be based on the distance between the samples and the number of samples in the neighborhood radius. In other words, samples that are slightly apart are similar. The spacing between samples is applied based on the attributes of the attribute from the feature selection stage. When two samples are slightly different from each other, the difference between these properties is a small amount; in fact, these two samples have similar properties to each other. So these two samples can be placed in a cluster. The basis of the DBSCAN clustering method is based on the fact that the number of similar samples more than one threshold can form a cluster together. Therefore, I have a threshold that should be adjusted optimally so that the process of education, accreditation, and testing of heart patients is done with high accuracy. This method has used adaptive thresholds in order to optimally differentiate between patient and healthy samples and to create patterns for diagnosis and prediction of new heart patients. Adaptive thresholds by searching the problem space try to find the best combination for both threshold values in order to diagnose heart patients.

In this section, first the prerequisites of the proposed method will be reviewed and then the details of the proposed method will be stated.

3.1. DBSCAN Clustering Algorithm. DBSCAN (density-based spatial clustering of applications with noise) is the first density-based clustering algorithm designed to collect data in desired shapes in the presence of noise in high-dimensional spatial and nonspatial data databases. The basic idea of DBSCAN is that for each object in a cluster, a local radius (Eps) is specified that must have a minimum number of objects (MinPts), meaning that the principle of local data points must exceed some thresholds. The neighborhood of the desired point "p" is defined as follows [25]:

$$N_{\text{Eps}} = \left\{ q \in \frac{D}{\text{dist}(p, q)} < \text{Eps} \right\}. \tag{1}$$

Here, D is the database of objects. If the neighborhood ε radius of a point P has a minimum number of points, this point is called the center point. The main point is defined as follows:

$$N_{\text{Eps}}(p) > \text{MinPts}. \tag{2}$$

Here, Eps and MinPts are user-defined parameters that mean the neighborhood radius and the minimum number of points in the neighborhood of a major point, respectively. If this condition is not met, this point will be considered a nonprincipal point. DBSACN searches for clusters by

examining the neighborhood of each object in the dataset. If the neighborhood of a p object is larger than MinPts, a new cluster is created with p as the primary object. It then frequently collects densely accessible objects from these main objects, which may involve merging a new cluster with new compression capabilities. This process ends when no new object is added to any cluster [25].

3.2. Proposed Method. This paper presents a density-based unsupervised approach to detect anomalies among heart patients. In this method, after selecting the basic features, the density-based clustering algorithm with adaptive parameters has been utilized to increase the clustering accuracy of normal samples and also to determine anomalous samples. The proposed approach first selects important features in the dataset that are more correlated with the class label of instructional examples. Features that do not play a significant role in patients' condition and cannot be useful for clustering should be excluded from other features. Such features are not correlated with the class label. Attribute correlation with a class label refers to features that can be used to determine the class label of instructional instances and those attributes that have little effect on determining the instance class label. Reducing such features allows the remaining features to represent the entire set of core data features in a better way. Thus, in addition to reducing the dimensions of the data, the prediction accuracy of the model also increases. In this method, we use a comparative approach to adjust the parameters of the clusters. In this approach, the size of the clusters is considered one of the main criteria for determining the neighborhood radius to join a point to the cluster. In fact, in the proposed method, it is assumed that the more points within the cluster, the less likely it is that the data related to that cluster would be anomalous, so a smaller neighborhood radius can be considered for it and the cluster can be more integrated and denser. Yet clusters with a small number of points inside are more likely to be anomalous and more distant from other clusters. Therefore, by increasing the neighborhood radius for such clusters, the density of these clusters can be reduced and the existing anomalies can be easily detected. Finally, samples that do not belong to any cluster are considered anomalies. Thus, equation (3) determines the desired neighborhood radius for each cluster in the proposed method.

$$\varepsilon = \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{\mu_{ij}}{n_j}, \tag{3}$$

where μ_{ij} is the average distance between clusters and n_j is the number of points within each cluster. Using this equation, the value of the neighborhood radius is determined as a factor of the number of points within the cluster and the size of the cluster. Determining the adaptive neighborhood radius will increase the distances between clusters and will detect existing anomalies. In the proposed method, another important parameter is that the number of connected points in the neighborhood radius is determined comparatively based on the cluster size. The hypothesis about the neighborhood radius also applies to the number of connected points within the cluster. In other words, clusters that are larger in size and have more points need fewer connected points to expand the cluster than clusters of smaller size and fewer points. Thus, the number of connected points is defined as a function of the amount of data within the cluster as

$$\text{Minpts} = \sum_{j=1}^{k} \frac{n_j}{N - n_j}, \tag{4}$$

where N is the total number of instances in the dataset. According to equation (4), the higher the number of points within a cluster, the minimum the need is for connected points in the neighborhood radius. Figure 1 illustrates the flowchart for the proposed method.

3.3. Evaluation Criteria for the Proposed Method. In this paper, we have used a labeled dataset and it is clear whether each sample is healthy or not. Therefore, in order to evaluate the proposed method, the sample testing process is done by dividing the data into a training dataset and test dataset. Test datasets are randomly selected part of the main dataset that the label of these samples in terms of heart disease is estimated by the proposed method. The proposed method provides an estimated class label for the data and compares it with the actual data label. Given that there are two classes of healthy people and people with heart disease in the present dataset, the confusion matrix can be utilized to compare samples. In this regard, the samples in the clusters and in the dataset that are part of the healthy samples are considered true positive samples, and the anomalous samples in the dataset, whose class label is patient, are considered true negatives, and these two groups will increase the accuracy of the proposed method. In contrast, samples that are clustered and whose class label is in the patient dataset are considered false positives, and anomalous samples whose class label is in the healthy dataset are considered false negatives. Thus, by obtaining the parameters related to comparison between the actual class and the predicted class from the dataset, evaluation criteria can be applied to the proposed method. There are several criteria for measuring the performance of clustering methods; one of the most famous is the F-measure criterion. This criterion helps to determine the accuracy of a clustering solution. F-measure, the measurement for a $\mathbf{C_j}$ cluster with respect to a particular class of $\mathbf{C_i}$, shows how a good $\mathbf{C_j}$ cluster describes the $\mathbf{C_i}$ class by calculating the harmonized average of the **precision** and **recall** parameters through the following equations:

$$p_{ij} = \frac{n_{ij}}{n_j}, \tag{5}$$

$$r_{ij} = \frac{n_{ij}}{n_i}, \tag{6}$$

$$f\left(C_i, C'_j\right) = \frac{2 * p_{ij} * r_{ij}}{p_{ij} + r_{ij}}, \tag{7}$$

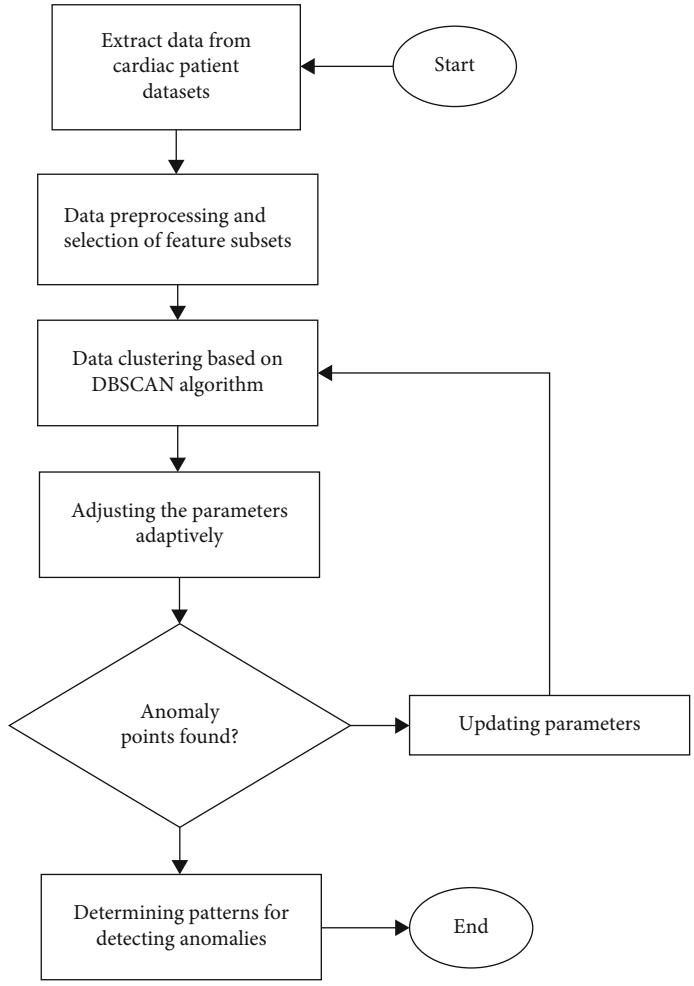

Figure 1: Flowchart of the proposed method.

where n_{ij} represents the number of specimens in the cluster C_i, which also appears in class C_j; n_i and n_j represent the number of specimens belonging to class C_i and class C_j, respectively; p_{ij} is the precision parameter of the fraction n_{ij} and n_j; r_{ij} is the recall parameter of the fraction n_{ij}; and n_i and n indicate the total number of samples in the dataset. In general, the F-measure is defined as follows:

$$F\text{-measure} = \sum_i \frac{n_{ij}}{n} \max_j f\left(C_i, C_j'\right). \tag{8}$$

A higher value for the F-measure indicates a better result for clustering. In case the amount of F-measure equals 1, it indicates that clustering is actually true.

4. Implementation of the Proposed Methods

This method has been implemented on a Corei7 laptop using MATLAB 2020 software. The standard DBSCAN clustering function to implement the proposed method and also the standard classification toolbox in MATLAB to implement the classification methods have been used. The dataset used in this article is from the Heart Disease Dataset located in the UCI standard data repository. The data in this dataset includes a list of 4 databases on the diagnosis of heart disease. All databases have similar attributes and properties that take continuous and numerical values. The datasets obtained from these databases have 76 raw and unprocessed properties. This number of features is too large, and not all of them can be used to identify anomalies in the data. Therefore, a number of attributes that are not related to the class tag should be removed, and only a number of these attributes that are useful should be extracted from this data.

4.1. Feature Selection. Feature selection (also known as attribute selection) occurs in a variety of areas, including machine learning, pattern recognition, data mining, and statistical analysis. In all of these areas, most of the objects studied contain irrelevant and redundant features in their descriptions which can significantly affect the analysis of data and consequently lead to having false results or even the creation of incorrect models.

Feature selection is the process of selecting the most useful features to build models in tasks such as classification, regression, or clustering. In addition, feature selection not only reduces the size of the data and facilitates its visualization and understanding but also usually leads to more

Table 1: An example of a feature set after the feature selection step.

Property	Allowed property value	Property type	Property	Allowed property value	Property type
Age	29-77	Numeric continuous	Maximum heart rate	70-202	Numeric continuous
Sex	0-1	Sequential	Exercise-related pain	0-1	Sequential
Type of chest pain	1-4	Exercise-induced depression	Sequential	0-6.2	Numeric continuous
Blood pressure	94-200	Numeric continuous	ST slope at the peak of exercise	1-3	Sequential
Cholesterol	126-564	Numeric continuous	The number of clogged main vessels	0-3	Sequential
Blood sugar	0-1	Sequential	Thalassemia patient status	3-7	Sequential
Electrocardiographic results	0-2	Sequential	Patient's cardiac status (predicted class)	0-4	Sequential

compact models with better generalizability. All of these features make feature selection an interesting area of research, where different feature selection methods have been introduced over the decades. In the present dataset, 76 raw features obtained from hospital databases are complex both in terms of volume and model training. Most of these features are redundant or useless and not only do not have a positive effect on the model for predicting heart anomalies but also can cause a loss of comprehensiveness and accuracy of the proposed method. Thus, the feature selection step is performed on this dataset and most of the existing features are removed through the filter feature selection method. The filtering method decides to delete or keep these features by considering the standard deviation of the data in the properties. In fact, when the standard deviation of the values of a property is less than a certain threshold, the values of this property are the same for both classes and this feature cannot play an effective role in determining the class label in healthy and patient instances, and also determining anomalies is impossible. Eliminating such features will help increase the accuracy of the proposed method.

In this method, a threshold value of 0.1 is used for standard deviation of features. Attributes with a standard deviation of less than 0.1 are removed. Obviously, the standard deviation as the distance between the value of the feature for the samples and the mean for the whole sample in that feature, if less than 0.1, indicates that the values of that feature are very close in both the patient and healthy patient classes. Therefore, determining a pattern for diagnosing heart patients using this feature is difficult and misleading. After the preprocessing step of feature selection, there are only 14 attributes left from the main dataset that satisfied the threshold. The explanation of the data and values in these 14 attributes is shown in Table 1.

As shown in Table 1, the remaining features are used as a representation of the main features in the proposed method. Also, according to Table 1, it can be seen that some of these properties are numerical and continuous and some are sequential. In the dataset used in the proposed method, a feature called cardiac status is predicted for each patient which is determined by specialist physicians and varies from 0 to 4. This feature is intended as a class tag for the training dataset, and model training occurs based on this tag. As it can be seen, instead of two labels in this dataset, there are 5 types of labels. The value of 0, which includes the majority of the samples, is shown for healthy people without any heart disease. Higher values, ranging from 1 to 4, are intended for different types of heart diseases for patients whose data have been stored.

4.2. Implementation of the Proposed Method. After the feature selection preprocessing step, the remaining dataset, which is classified into 5 predicted classes, is divided into two classes of healthy and sick people, and all types of heart diseases are considered one class. This is done to detect anomalies in each class. Since in the proposed method, the number of samples showing a class of heart disease is small and after clustering, the whole existing cluster may be identified as an anomaly, to avoid this phenomenon, the whole heart samples of diseases are put in one class, and based on two classes of healthy people and sick patients with various heart diseases, a decision is made for an anomaly in a sample. In the proposed method, each sample is first divided into two parts: training and test datasets, and model training is created on 1000 training samples, and this data is entered into the model. The proposed model, considering the innovation of using the neighborhood radius and the number of connected neighbors in the adaptive neighborhood radius, determines the optimal values of the density-based clustering parameters in each step of the clustering. Thus, the number of clusters and the number of samples within each cluster may change at each step of clustering, and eventually the proposed method will converge towards the optimal points for the parameters. Figure 2 shows the clustering steps in the proposed method.

As shown in Figure 2, the clustering steps based on adaptive density can be seen in the proposed method. In the first step, the values of the neighborhood radius parameters and the number of connected neighbors in the neighborhood radius are both considered equal to zero, and the number of clusters found is approximately equal to the number of samples in the dataset. These values are then updated, and by increasing the values for the parameters, they approach the optimal point and decrease the number of clusters. As

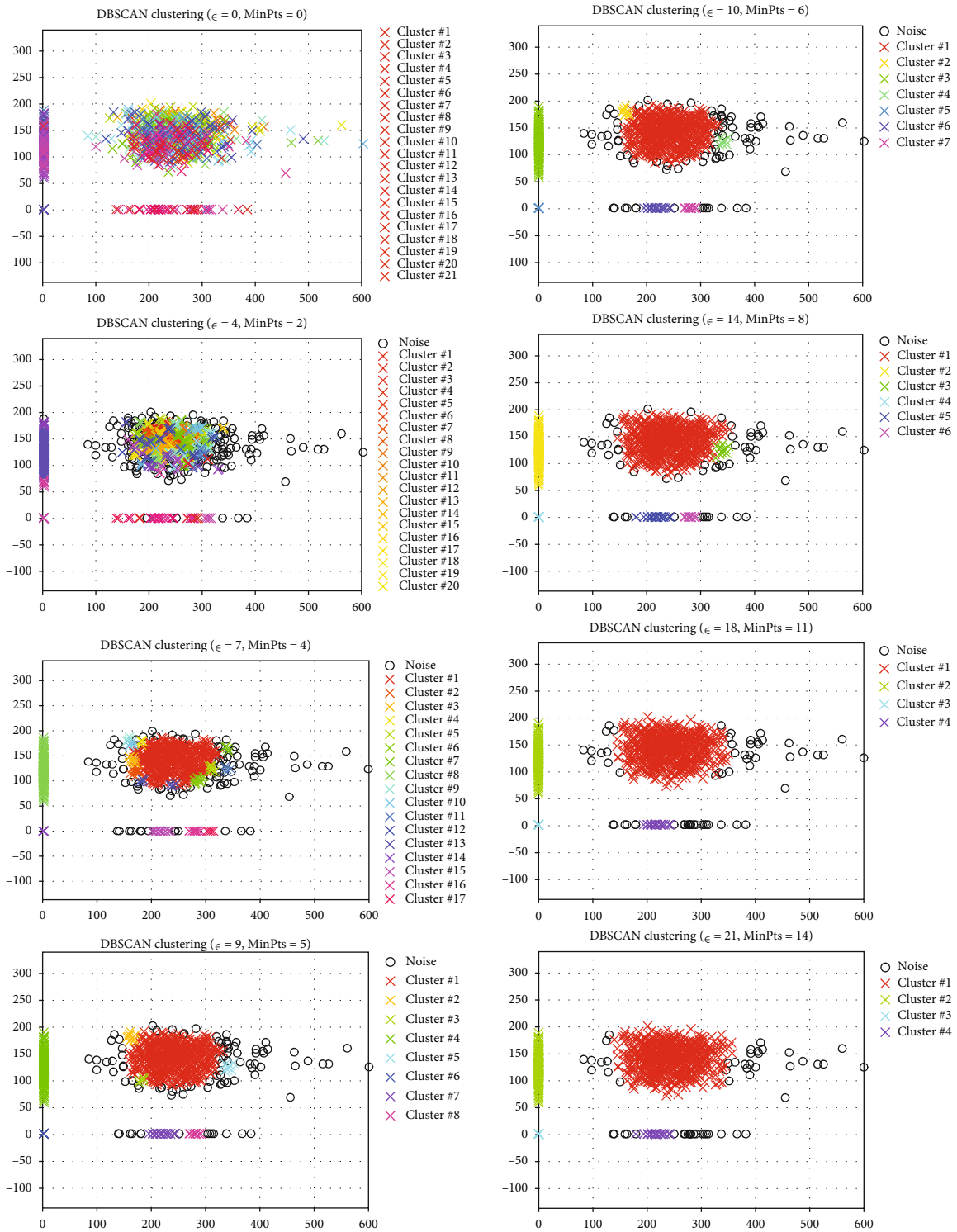

Figure 2: Clustering steps in the proposed method.

shown in Figure 2, the threshold values $[0, 0]$ for both thresholds were initially selected, and by examining the possible cases, we finally reached the best values for both thresholds. At each stage of the Huff threshold value change, clustering accuracy was evaluated to diagnose heart patients. Finally, in the tenth step of the optimal point, parameters are found on the values of the neighborhood radius equal to 21 and the number of connected neighbors in the neighborhood radius equal to 14 optimal points. At this point, the model error has reached its minimum value and the model accuracy shows a high value. Figure 2 shows clusters with colored crosses and anomalous specimens with black circular dots which are marked as noise. These samples do not belong to any of the clusters and have been identified as

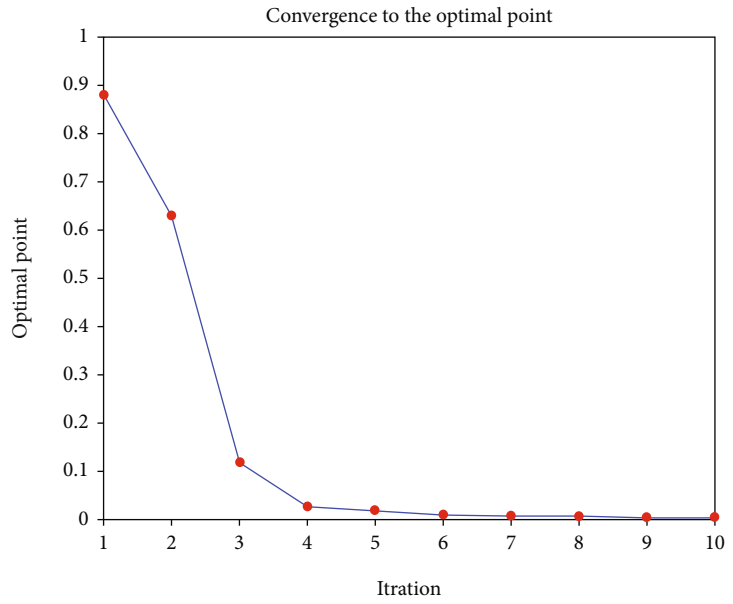

Figure 3: Convergence of clustering parameters towards the optimal point.

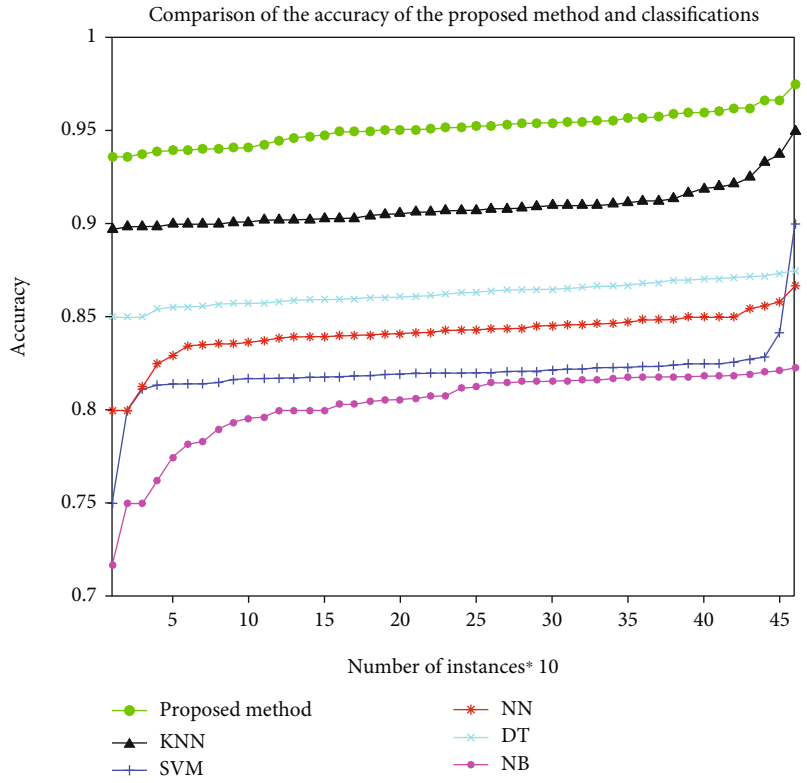

Figure 4: Diagram of classification accuracy of the proposed method for test data.

anomalies. In the next section, we will evaluate these anomalies based on the class label provided by physicians to determine the ability of the proposed method to find the sample in order to evaluate new heart diseases. Figure 3 shows the convergence parameters of the clustering parameters towards the optimal point.

As shown in Figure 3, the clustering parameters have a very steep slope in convergence according to relations (3) and (4) presented for the neighborhood radius and the number of connected points in the neighborhood radius in the proposed method, and the optimal points were found after ten times repetition. According to the convergence diagram

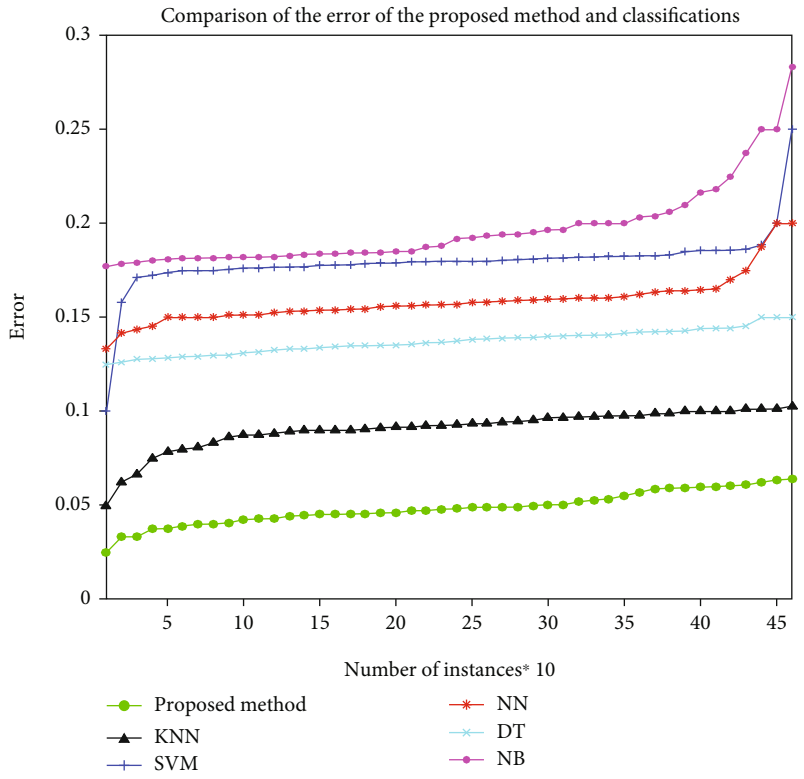

Figure 5: Diagram of the proposed method error for test data.

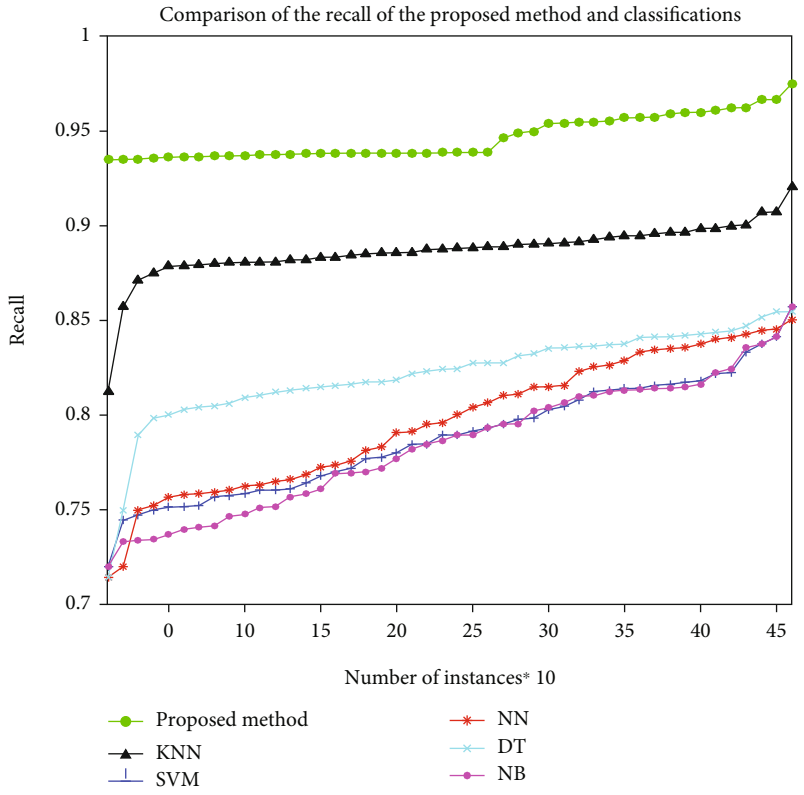

Figure 6: Sensitivity diagram of the proposed method for test samples.

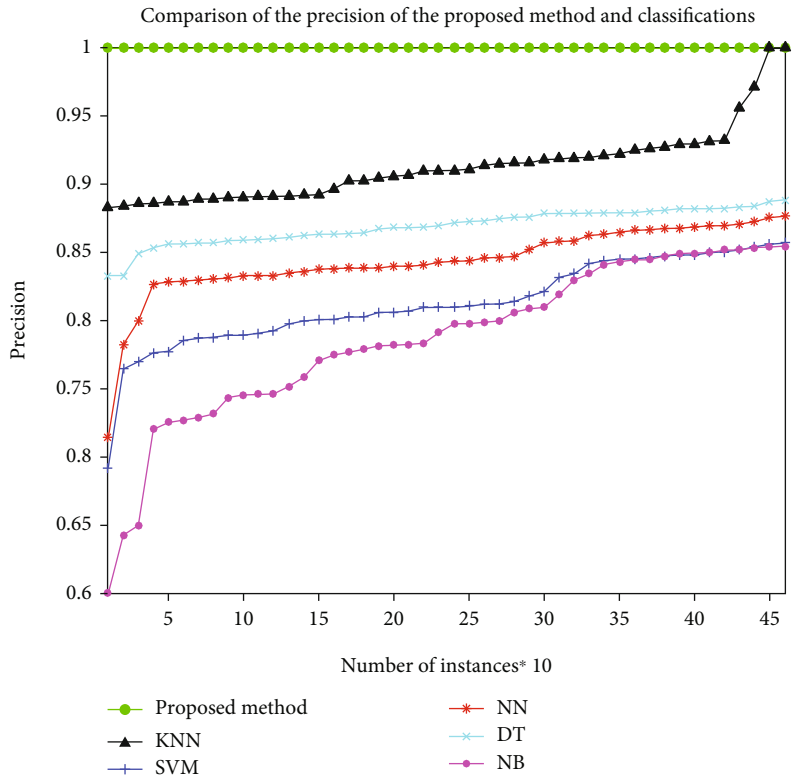

Figure 7: Diagram of the accuracy of the proposed method for test samples.

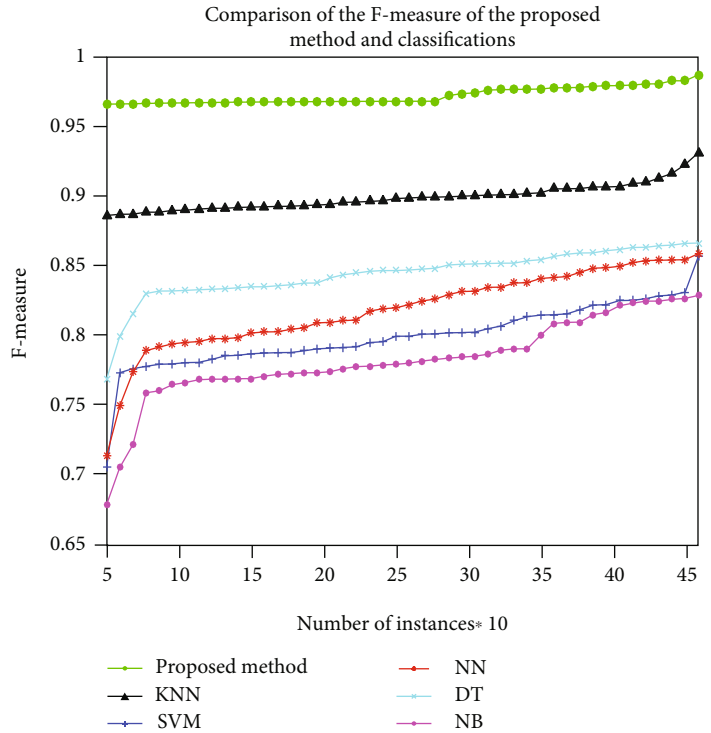

Figure 8: *F*-measure diagram of the proposed method for test samples.

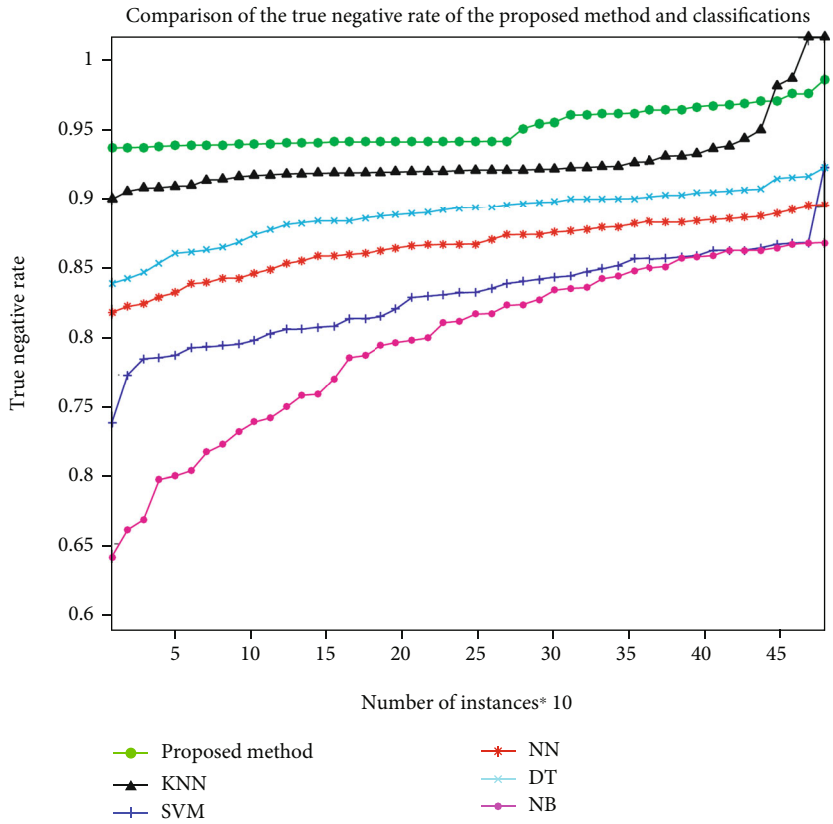

Figure 9: True negative rate diagram of the proposed method for test samples.

Table 2: Values related to evaluation criteria.

Criterion	*F*-measure	Precision	Recall	True negative rate	Accuracy
Proposed method	97.25	99.99	94.65	93.57	95.14
KNN	89.95	91.33	88.68	92.75	90.92
SVM	79.99	81.32	78.82	94.63	82.04
NN	81.93	84.4	79.64	87.6	84.12
DT	84.47	86.94	82.16	89.5	86.28
NB	78.35	78.67	78.43	81.91	80.29

in Figure 3, it can be said that the proposed method has a high ability to find the parameters related to each cluster according to the cluster size, and cluster size plays an essential role in determining anomalies.

4.3. Evaluation of the Proposed Method. After implementing the proposed method on the cardiac patient dataset which has been reduced based on preprocessing and feature selection, we will evaluate the proposed method based on the analogy of anomalies obtained by the proposed method with the label provided for patients by doctors. Anomalous specimens are specimens whose behavior is very different from that of healthy specimens, and the discovery of these specimens could lead to the discovery of possible new heart disease specimens and people at risk for heart disease. Thus, in the proposed method, in addition to identifying current patients with heart disease, possible cases of heart disease in the form of anomalies are also identified. In order to evaluate the proposed method, true positive, false positive, true negative, and false negative samples were used. The high accuracy of the proposed method indicates the high ability of this method in teaching the model based on selected features during the feature selection step and matching the parameters related to density-based clustering and anomalies discovered in the proposed method. The accuracy diagram of the proposed method for 470 test samples is shown in Figure 4.

As it is indicated in Figure 4, the proposed method performed better than other classification methods in terms of accuracy by detecting anomalous samples in the heart patient database. The proposed method, relying on the adaptation of clustering parameters based on cluster size, has been able to find a safe range for the relevant cluster and class, and samples that are outside this range have been

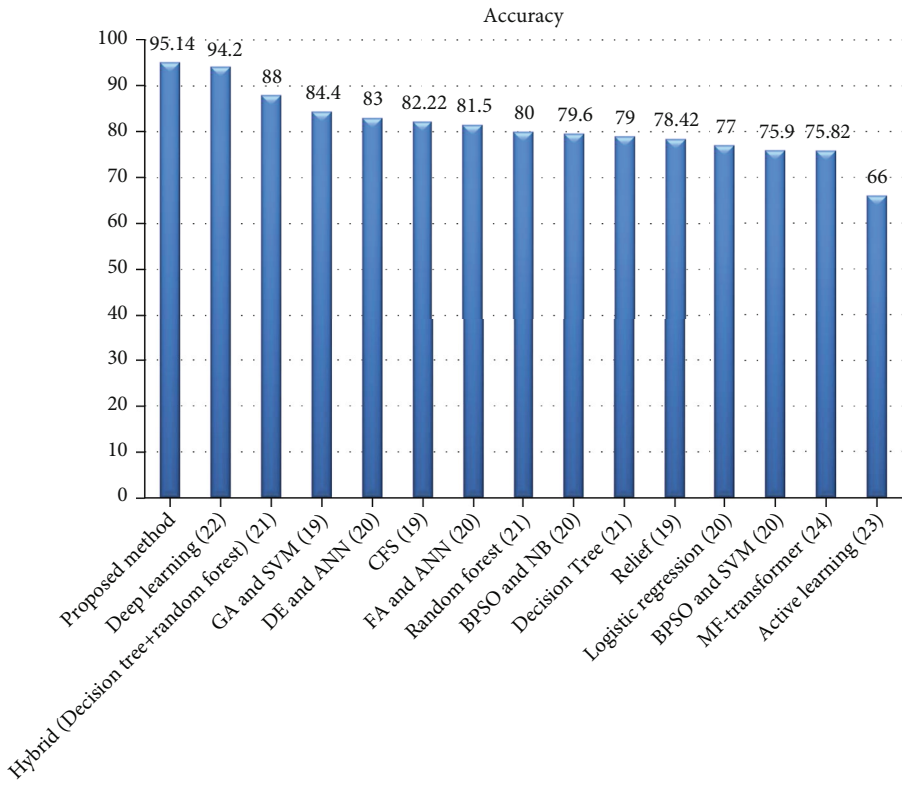

Figure 10: Comparison of the proposed method with previous methods.

identified as anomalous samples. The identified anomaly samples, according to the class label predicted by doctors, are mostly related to heart patients, and this phenomenon has increased the accuracy of the proposed method. We can now discuss the error range in the proposed method. Figure 5 shows the error diagram of the proposed method.

As shown in Figure 5, the error of the proposed method versus accuracy is very small compared to other classification methods; this amount reaches 5% of the total test data. Another criterion that has been measured in the proposed method is the sensitivity criterion which is defined as the ratio of anomalous samples detected whose true class is heart disease to the total anomalies detected in the proposed method. Figure 6 shows the graph related to the sensitivity criterion in the proposed method.

As shown in Figure 6, more than 92% of the detected anomalies belong to the class of heart patients and this indicates the high ability of the proposed method to detect anomalies in the proposed method compared to other classification methods. Another criterion used in the proposed method is accuracy which is clustered as a percentage of cardiac patients and the anomalies detected belonging to the cardiac class. Figure 7 shows the correctness of the proposed method.

As shown in Figure 7, most of the patient specimens and discovered anomalies that have been clustered belong to the cardiac patient class, indicating the accuracy of clustering in the proposed method. Another criterion used in the proposed method is F-measure which is a combination of two criteria, accuracy and sensitivity. Figure 8 shows the F-measure diagram of the proposed method.

As shown in Figure 8, the F-measure which is a combination of accuracy and sensitivity has a high value of approximately 96%. Figure 9 shows the true negative rate diagram of the proposed method for the test data.

As shown in Figure 9, healthy individuals present in the dataset were detected more accurately based on the proposed method, compared to other classification methods. Accordingly, based on the presented graphs, Table 2 shows the average values to the evaluation criteria.

As presented in Table 2, the proposed method has better performance in terms of evaluation criteria compared to other classification methods. The proposed method, by carefully selecting useful features in the diagnosis of heart disease, provides suitable data for ready-to-die processing. It is obvious that the more accurate the input to machine learning methods, the more accurate the training and the better the results. Also, the proposed method solves the problems in data clustering and adapting the parameters in the DBSCAN clustering method, and well-constructed clusters have been obtained. The main cluster represents healthy instances, and the other clusters represent anomalies and diseased instances. The accuracy of predicting test samples in the proposed method is higher than that in other classification methods. This is because the proposed method carefully selects input features. The scratching method has also been manipulated to more accurately separate healthy specimens from heart patients. The existence of well-constructed

clusters with distinction between healthy samples and heart patients is a proof of the accuracy of feature selection and the effect of adaptive parameters on the DBSCAN clustering method.

4.4. Comparison of the Proposed Method with Previous Methods. After evaluating the proposed method, in order to measure the validity of the performance of the proposed method, we compare it with previous methods in this field. As shown in Table 2, the performance accuracy of the proposed method was determined on the cardiac patient dataset and related diagrams and evaluation criteria and the proposed method was compared with other classifiers. In this section, we compare the proposed method with the methods that have made improvements to the classifications in order to increase the accuracy of heart patients' predictions. Based on this, the proposed method can be compared with previous methods [19–24] on the same dataset. Therefore, Figure 10 shows a comparison of the proposed method with previous methods in predicting the class label of heart patients.

As shown in Figure 10, the proposed method has a higher accuracy in diagnosing cardiac patients compared with other previous methods owing to the adaptation of density-based clustering parameters to the number of samples in the cluster and the average intercluster distances.

5. Conclusion

Heart disease prediction systems are absolutely useful as an aid to maintain and monitor the health of the patient community, thereby reducing mortality even in young and middle-aged people. The heart disease prediction system can be an important asset, and in this way, it helps ordinary people to be aware of their health status since they are usually nonchalant about their periodic health examinations and they do not have the necessary knowledge regarding their health status. With the results of the heart disease prediction system in hand, if people have heart disease, they can be aware of the development of the disease and prevent the progression of the disease in the early stages. Then, based on the various parameters compared, appropriate medical suggestions can be provided for patients. The accuracy of diagnosis of heart disease prediction systems depends on the accuracy of the proposed model in determining diagnostic patterns. Therefore, the more accurate the model, the greater the ability to predict patients at risk for heart disease. In this study, a density-based unsupervised approach is proposed to detect anomalies in heart patients. In this study, after extracting the basic features, the density-based clustering method (DBSCAN) with adaptive parameters has been used to increase the clustering accuracy of normal samples and determine anomalous samples. The experimental results show that the proposed method has a good performance in terms of evaluation criteria. The high accuracy of the proposed method indicates the high ability of this method in training the model based on the selected features during the feature selection step and matching the parameters related to density-based clustering and anomalies discovered in the proposed method. Therefore, it can be said that the proposed method has a higher accuracy in diagnosing heart patients compared to other previous methods due to the adaptation of density-based clustering parameters to the number of samples in the cluster and the average intercluster distances. In order to make suggestions for future work, it can be said that since a lot of effort has been made in predicting people with heart disease, it is possible to use a combination of supervised methods such as decision tree, K-nearest neighbor and Bayesian networks or unsupervised methods such as clustering, with feature selection based on metaheuristic methods to increase the accuracy of classification and prediction of patients with this particular type of disease. It is also possible to increase the accuracy of predicting disease progression in patients with heart disease by selecting the used optimization criteria.

The use of feature selection approaches based on metaheuristic methods can be suggested as a continuation of this research in the future. Metaheuristic methods take the features as input and build the initial population according to the features in the dataset and try to find the smallest subset of features with the least amount of error in sample classification and predicting heart patients. By increasing the accuracy of input data to machine learning methods, we can expect that learning models identify an accurate pattern for diagnosing and predicting heart disease.

Data Availability

The data used to support the findings of this study are available from the authors' upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] B. Nedelcu, V. Sgârciu, and A. Jigau, "Mining medical data," *2018 10th international conference on electronics, computers and artificial intelligence (ECAI)*, 2018IEEE, 2018.

[2] V. Kotu and B. Deshpande, *Data Science: Concepts and Practice*, Morgan Kaufmann, 2018.

[3] S. P. Shaji, "Predictionand diagnosis of heart disease patients using data mining technique," *2019 international conference on communication and signal processing (ICCSP)*, 2019IEEE, 2019.

[4] A. Aldallal and A. A. A. Al-Moosa, "Using data mining techniques to predict diabetes and heart diseases," *2018 4th international conference on Frontiers of signal processing (ICFSP)*, 2018IEEE, 2018.

[5] C. Bou Rjeily, G. Badr, A. Hajjarm El Hassani, and E. Andres, "Medical data mining for heart diseases and the future of sequential mining in medical field," in *Machine Learning Paradigms*, pp. 71–99, Springer, 2019.

[6] P. Singh, S. Singh, and G. S. Pandi-Jain, "Effective heart disease prediction system using data mining techniques," *International Journal of Nanomedicine*, vol. Volume 13, pp. 121–124, 2018.

[7] R. Fadnavis, K. Dhore, D. Gupta, J. Waghmare, and D. Kosankar, "Heart disease prediction using data mining," *IOP Publishing.*, vol. 1913, no. 1, p. 012099.

[8] C. Bou Rjeily, G. Badr, A. Hajjarm El Hassani, and E. Andres, "A comprehensive looks at data mining techniques contributing to medical data growth: a survey of researcher reviews," in *Recent Developments in Intelligent Computing, Communication and Devices*, pp. 21–26, Springer, 2019.

[9] A. Singh and R. Kumar, "Heart disease prediction using machine learning algorithms," *2020 international conference on electrical and electronics engineering (ICE3)*, 2020IEEE, 2020.

[10] Y. Liu, Z. Li, C. Zhou et al., "Generative adversarial active learning for unsupervised outlier detection," *IEEE Transactions on Knowledge and Data Engineering*, p. 1, 2019.

[11] R. A. R. Ashfaq, X. Z. Wang, J. Z. Huang, H. Abbas, and Y. L. He, "Fuzziness based semi-supervised learning approach for intrusion detection system," *Information Sciences*, vol. 378, pp. 484–497, 2017.

[12] R. G. Franklin and B. Muthukumar, "Survey of heart disease prediction and identification using machine learning approaches," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 2020IEEE, 2020.

[13] "Heart Disease Prediction," http://archive.ics.uci.edu/ml/datasets/statlog+(heart). 2022.

[14] R. Singh and E. Rajesh, "Prediction of heart disease by clustering and classification techniques," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 5, pp. 861–866, 2019.

[15] Y. Liu, D. Liu, F. Yu, and Z. Ma, "A novel local density hierarchical clustering algorithm based on reverse nearest neighbors," *Mathematical Problems in Engineering*, vol. 2019, 10 pages, 2019.

[16] J. L. Lima, D. Macêdo, and C. Zanchettin, "Heartbeat anomaly detection using adversarial oversampling," *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019IEEE, 2019.

[17] E. Talab, O. Mohamed, L. Begum, F. Aloul, and A. Sagahyroon, "Detecting heart anomalies using mobile phones and machine learning," *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2019IEEE, 2019.

[18] P. Umasankar and V. Thiagarasu, "Decision support system for heart disease diagnosis using interval vague set and fuzzy association rule mining," *2018 4th International Conference on Devices, Circuits and Systems (ICDCS)*, 2018IEEE, 2018.

[19] C. B. Gokulnath and S. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease," *Cluster Computing*, vol. 22, no. S6, pp. 14777–14787, 2019.

[20] T. Vivekanandan and N. C. S. N. Iyengar, "Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease," *Computers in Biology and Medicine*, vol. 90, pp. 125–136, 2017.

[21] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 2021IEEE, 2021.

[22] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Computational Intelligence and Neuroscience*, vol. 2021, 11 pages, 2021.

[23] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, "Multi-label active learning-based machine learning model for heart disease prediction," *Sensors*, vol. 22, no. 3, p. 1184, 2022.

[24] J. Qiu, J. Zhu, M. Rosenberg, E. Liu, and D. Zhao, "Optimal transport based data augmentation for heart disease diagnosis and prediction," https://arxiv.org/abs/2202.00567, 2022.

[25] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "DBSCAN: past, present and future," *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, 2014IEEE, 2014.