SN Computer Science (2021) 2:112
https://doi.org/10.1007/s42979-021-00518-7

ORIGINAL RESEARCH

A Data-Driven Heart Disease Prediction Model Through *K*-Means Clustering-Based Anomaly Detection

Rony Chowdhury Ripan[1] · Iqbal H. Sarker[1] · Syed Md. Minhaz Hossain[1,2] · Md. Musfique Anwar[3] · Raza Nowrozy[4] · Mohammed Moshiul Hoque[1] · Md. Hasan Furhad[5]

Received: 13 August 2020 / Accepted: 11 February 2021 / Published online: 24 February 2021

Abstract

Heart disease, alternatively known as cardiovascular disease, is the primary basis of death worldwide over the past few decades. To make an early diagnosis, a data-driven prediction model considering the associate risk factors in heart disease can play a significant role in healthcare domain. However, to build such an effective model based on machine learning techniques, the *quality of the data*, e.g., data without "anomalies" or outliers, is important. This research investigates *anomaly detection* in the healthcare domain to effectively predict heart disease using unsupervised *K-means clustering* algorithm. Our proposed model first determines an *optimal* value of *K* using the Silhouette method to form the clusters for finding the anomalies. After that, we eliminate the identified anomalies from the data and employ the five most popular machine learning classification techniques, such as *K*-nearest neighbor, random forest, support vector machine, naive Bayes, and logistic regression to build the resultant prediction model. The efficacy of the proposed methodology is justified using a standard heart disease dataset. We also take into account the data plotting to test the exactness of the detection of anomalies in our experimental analysis.

Keywords Anomaly detection · Healthcare · *K*-means clustering · Heart disease prediction

Introduction

The new digital world is overwhelmed by an enormous volume of data. Most organizations have no problem capturing an ample amount of data. However, the challenging task for them is to elucidate and extract the required meaningful knowledge or information from these vast data with the help of data science and machine learning methods [28, 30].

Several data mining techniques are used to find interesting and meaningful relationships among data. One such technique is called anomaly detection. The data characteristics, which are different from normal behaviors, are called anomalies. In some cases, such anomalies or outliers are considered as noise that affects the prediction model [26]. Detection of anomalies has recently occupied an overwhelming research interest owing to its necessity in various domains to get critical actionable information from large datasets.

For example, an abnormal MRI image may indicate the presence of malignant tumors [32]. It is widespread to observe abnormal traffic patterns in a computer network that sends out sensitive data to an unauthorized destination [14]. One may notice unusual readings from a spacecraft sensor could signify a fault in some components of the spacecraft [11]. In the medical and public health areas, anomaly detection usually operates with patient information. Due to different causes, such as a rare patient condition or instrumentation failures, or recording errors, the data may have anomalies. These anomalies can lead to misclassification or bad classification for a classification model and give health practitioners some hard time. So, it is important to detect anomalies in the heart disease dataset for better heart disease classification.

Recently, some studies have focused on anomaly detection for larger datasets in [9, 20, 27, 35, 40]. An elaborate

✉ Iqbal H. Sarker
iqbal.sarker.cse@gmail.com

[1] Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chittagong 4349, Bangladesh

[2] Premier University, Chittagong, Bangladesh

[3] Jahangirnagar University, Dhaka, Bangladesh

[4] Victoria University, Footscray, VIC 3011, Australia

[5] Canberra Institute of Technology, Canberra, Australia

discussion on these methodologies are given in “Related Work”. The common phenomenon of the most existing techniques is to construct a profile of normal instances which is the challenging part to find a sufficient number of normal profiles.

In this work, we propose a model of *detecting anomalies* in the heart disease data based on the *K-means clustering* algorithm to improve heart disease classification that extends and revises our earlier version [22]. The major benefit of our methodology is that it does not require the creation of normal profiles or experience of prior anomaly information in the heart disease training dataset. Before applying K-means clustering, first, the optimal value of K is measured using the Silhouette method. After that, K-means clustering is applied, and anomalies are detected using some threshold values according to their respective cluster. Usually, the anomalous instances are located in sparse or small clusters and far from their respective cluster’s centroids. Anomalous instances are removed, and the exactness of anomalies is then tested using the boxplot method and scatter plot before and after removing anomalies. Besides, five popular machine learning classification techniques [31] such as K-nearest neighbor (KNN), random forest (RF), support vector machines (SVM), naive-Bayes (NB), and logistic regression (LR) are applied to measure the effectiveness of the proposed methodology in anomaly detection to predict heart disease. We also present the effect of different cluster values of K on the efficacy of the proposed method.

The remainder of this paper is organized as follows: in the next section, we discuss the related works about anomaly detection and heart disease prediction followed by which details of the suggested anomaly detection model are given. The two subsequent sections show the experimental results and discussions on those results, respectively. Finally, we conclude the paper by providing the future directions of this research.

Related Work

Existing research works on anomaly detection can be categorized as supervised, semi-supervised, and unsupervised classification techniques.

Supervised system for anomaly detection has several steps: (i) instances are first hand-labeled as normal or abnormal, (ii) the dataset is divided into training and testing set, and (iii) the system then extracts a wide variety of features to train the classifier to predict the classification of a test data. In some applications, the experts may have to label the normal objects manually, and any other objects that do not match the normal objects are reported as anomalies. Examples of such models are Bayesian networks and Support vector machines [34]. Janakiram et al. [13] proposed an anomaly detection method based on Bayesian belief networks for detecting the anomalies using conditional dependencies among the observations of the attributes in the sensor streamed data. Steven Mascaro et al. [16] detects anomalies using Bayesian networks. Supplementary data-producing dynamic and static Bayesian network model learns anomalies from real-world Automated Identification System (AIS) data. The model proposed by Xu et al. [37] detects anomalies using Continuous-Time Bayesian Networks (CTBN). The CTBN avoids a fixed update interval to create a continuous-time model. They build generative models from the normal training data, and anomalies are detected based on their likelihood under this norm. Mohamed and Kavitha [17] proposed an anomaly detection method using Support vector machines in the wireless sensor networks. Their proposed model classifies the wireless sensor node data as a network anomaly or cluster anomaly, or local anomaly. The major drawback of all these previous supervised approaches is that they require many manually labeled instances, which are expensive to produce and sometimes limited in quantity.

The semi-supervised anomaly detection approaches learn from a small number of labeled training instances, then detect anomalies from a large amount of unlabeled training data. Examples of such models are spacecraft fault detection [11], fuzzy rough semi-supervised anomaly detection [38]. Fujimaki et al. [11] proposed a novel anomaly detection method for spacecraft based on Kernel Feature Space and directional distribution. Their approach detects anomalies in the spacecraft system by changing the model’s behavior from the previous telemetry data and incoming data. Xue et al. [38] presented a fuzzy rough semi-supervised anomaly detection (FRSSOD) approach, in which they detect anomalies with the help of some labeled samples and fuzzy rough C-means clustering. This method introduces an objective function for minimizing the sum squared error of clustering results and the variation from previously known labeled examples and the number of anomalies. Here a center, a lower estimation, and a fuzzy boundary represent a cluster. The data points located in this boundary can be anomalies. Hence, creating an anomaly detection model based on few labeled anomalies is unlikely to be effective.

Unsupervised anomaly detection models have significant benefits since it does not require any labeled training data. These techniques make an implicit assumption that anomaly instances are less more frequent than normal instances. For detecting a novel event, the unsupervised mode can be better and recalibrate its definition of normal [33]. Example of such model is Isolation Forest [15], DBSCAN [4], K-means clustering algorithm. Sun et al. [33] proposed an approach of unsupervised anomaly detection using an extended Isolation forest algorithm. They applied their approach to an enterprise dataset and isolated anomaly instances from the baseline user model using a single feature or combined features. A novel anomaly detection system based on Isolation Forest was proposed by Ding and Fei [8], in which they used the sliding window frame, taking into account the drift phenomenon principle. An unsupervised anomaly detection model using the DBSCAN algorithm was suggested by Ranjith et al. [21]. They sought to figure out anomalies from a traffic dataset, in which a path is said to be an anomaly if it does not match with the trained model. Münz et al. [19] also tried to find anomalies in a traffic dataset using the *K*-means clustering algorithm. Yoon et al. [39] proposed an approach to detect anomalies in software measurement data using the *K*-means clustering algorithm.

There are several studies on heart disease prediction. A comparative research on the prediction of coronary artery heart disease was performed by Ayon et al. [2], which implemented multiple data mining techniques such as logistic regression (LR), support vector machine (SVM), deep neural network (DNN), decision tree (DT), naive Bayes (NB), random forest (RF), and *K*-nearest neighbor (KNN). Mohan et al. [18] propose a novel method in which they aim to improve the prediction accuracy by finding significant features. The prediction model is introduced with different combinations of features, and several known classification techniques are applied to evaluate the performance of the model. Dessai et al. [7] presented an efficient approach for heart disease prediction based on a probabilistic neural network. However, the approaches mentioned above did not consider the impact of anomaly instances on the evaluation results.

Our suggested model of anomaly detection is based on an unsupervised method in which the optimal *K*-means clustering algorithm is used to cluster anomalies in the data for heart disease. Using the Silhouette approach, the optimum cluster value of *K* has been calculated, and classification techniques are used to predict heart disease by removing anomalies effectively.

Methodology

This section presents the details of our proposed anomaly detection model, which has five different modules. At first, we have a data preparation module to handle missing values in heart disease data. Next, in the clustering module, our model first determines the optimal value of *K* using the Silhouette method and then, apply *K*-means clustering on the heart disease dataset using the obtained optimal *K* value. Then, anomalies have been detected in the anomaly detection module using max threshold (MaxT) and min threshold (MinT). After that, anomalies are removed in the anomaly removal module. Finally, the effectiveness of the proposed model has been justified by five classifiers such as *K*-nearest neighbor (KNN), random forest (RF), support vector machines (SVM), naive-Bayes (NB), and logistic regression (LR) in the prediction module. A graphical diagram of anomaly detection in heart disease data is shown in Fig. 1.

Data Preparation Module

Usually, instances of healthcare datasets have many healthcare characteristics and associated information that can be used to create a model for anomaly detection. In this work, we use a benchmark heart disease dataset in Kaggle [23]. This dataset contains 303 instances, and each instance has 13 features. Table 1 lists the features with descriptions.

Real-world datasets often contain noisy instances as well as an enormous number of attributes, null values, or missing values, inconsistency data because of their origin from heterogeneous sources [26]. To improve the quality of the datasets and the performance of subsequent steps, we need to apply several data preprocessing techniques to clean noisy data, replace missing values, correct inconsistencies in data, and eliminate redundant features. In our dataset, there are two attributes ("chol" and "thalach") that have missing values, as shown in Fig. 2. There are several methods to handle missing values in the dataset. Since our heart disease dataset has only two attributes with missing values and the number of missing values is small, we used the imputation strategy. There are many ways to impute missing values. Since these

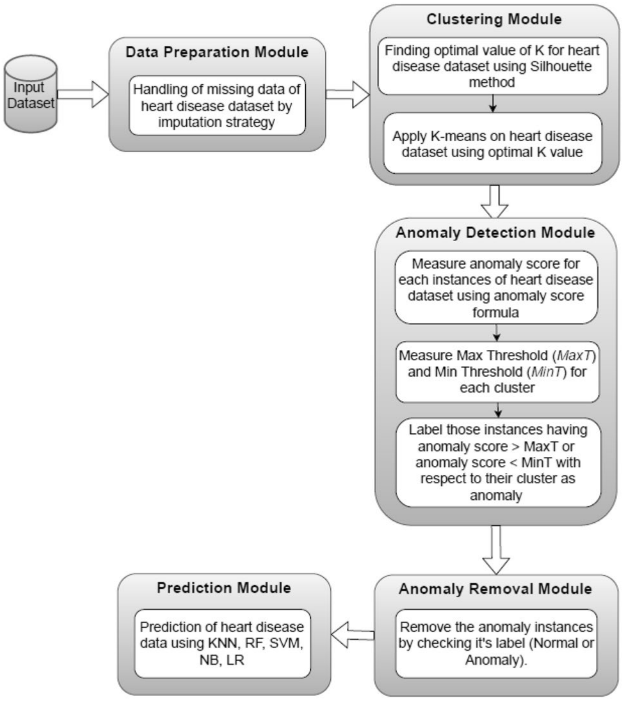
Input Dataset
Data Preparation Module
Handling of missing data of heart disease dataset by imputation strategy
Clustering Module
Finding optimal value of K for heart disease dataset using Silhouette method
Apply K-means on heart disease dataset using optimal K value
Anomaly Detection Module
Measure anomaly score for each instances of heart disease dataset using anomaly score formula
Measure Max Threshold (MaxT) and Min Threshold (MinT) for each cluster
Label those instances having anomaly score > MaxT or anomaly score < MinT with respect to their cluster as anomaly
Anomaly Removal Module
Remove the anomaly instances by checking it's label (Normal or Anomaly).
Prediction Module
Prediction of heart disease data using KNN, RF, SVM, NB, LR

Fig. 1 The proposed data-driven heart disease prediction model using machine learning techniques

Table 1 Features of heart disease dataset

Feature's name	Datatype	Feature's description
Age	Integer	Age in years
Sex	Integer	1 = male; 0 = female
cp	Integer	Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
trestbps	Integer	Resting blood pressure (in mmHg)
chol	Integer	Serum cholesterol (in mg/dl)
fbs	Integer	Fasting blood sugar ≥ 120 mg/dl (1 = true; 0 = false)
restecg	Integer	Resting electrocardiographic results (0 = normal; 1 = having ST–T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	Integer	Maximum heart rate achieved
exang	Integer	Exercise induced angina (1 = yes; 0 = no)
oldpeak	Float	ST depression induced by exercise relative to rest
slope	Integer	The slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3: downsloping)
ca	Integer	Number of major vessels (0–3) colored by fluoroscopy
thal	Integer	3 = normal; 6 = fixed defect; 7 = reversible defect

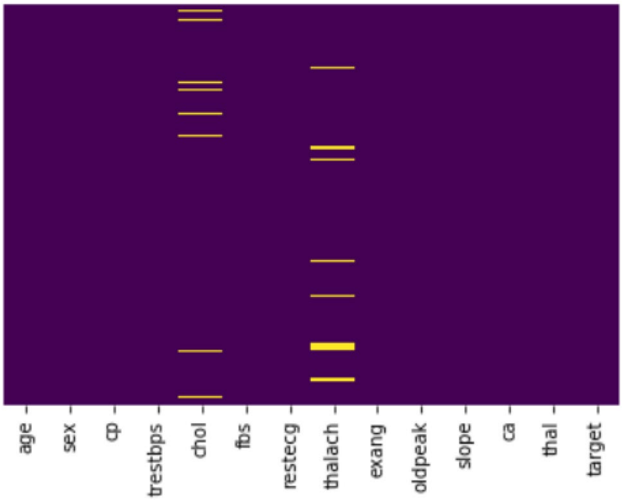

Fig. 2 Representation of attributes with null values using heatmap

two attributes only contain numeric values, we impute the mean value for each attribute.

Clustering Module

Clustering is the method of grouping data points or objects, or instances into multiple clusters. Data instances within a cluster have higher similarity but have dissimilarities to instances in another cluster [12]. Similarities and dissimilarities are measured on the attribute values narrating the data instances and often involve distance measures. On the same dataset, different clustering techniques may provide different clustering results. Clustering techniques have been widely used in many applications in detecting anomalies mentioned above in "Related Work". We choose to apply *K*-means clustering to detect the anomalies in heart disease data.

K-Means Clustering

The *K*-means algorithm [10] is an unsupervised clustering algorithm. It takes the number of clusters and the dataset as input and provides the result as a cluster collection. The *K*-means algorithm defines the mean value of the instances within the cluster as the center of a cluster. Next, *K* is chosen randomly from the instances representing a center or cluster in the dataset. An instance is allocated to the closest cluster for each of the residual instances, depending on the Euclidean distance between the mean of the cluster and the instance. It then continuously optimizes the positions of the centers for all the clusters. For each cluster, using the instances allocated to that cluster in the previous iteration, the new mean is determined as the center. Then all the instances of each cluster are reallocated to the modified means.

The iterations continue until no more repositioning is required for the centers of all clusters. Different *K* values can lead to various outcomes. So it is important that the optimal value of *K* is found. In this research, we used the Silhouette approach to determine the optimal value of *K* from the heart disease dataset. The silhouette method calculates the silhouette score for a range of *K* values, i.e. the silhouette coefficient for all instances. The Silhouette Coefficient for each instance is calculated using Eq. 1 [24]. In this equation, *a* represents the average distance between the instances within-cluster, and *b* represents the average distance between the instance and the nearest cluster or clusters. The value of the silhouette coefficient varies from −1 to +1, where +1 means the best fit for the cluster, and −1 means the worst fit for the cluster.

$$\text{Silhouette coefficient} = \frac{(b-a)}{\max(a,b)}. \tag{1}$$

Anomaly Detection Module

Clustering-based methods can detect anomalies by analyzing the relationship between instances and clusters. If there is a huge distance between the instance and the clusters, it can be called an anomaly. In our approach, we calculate the anomaly score of an instance (as shown in Eq. 2) based on the distance between the instance and the center of its nearest cluster [12].

$$\text{Anomaly score} = \frac{\text{distance}(o, C_0)}{L}. \tag{2}$$

In this formula, $\text{distance}(o, C_0)$ represents the distance between instance o and cluster center C_0 and L represents the average distance of that cluster. So Anomaly score in Eq. 2 calculates the ratio of each instance's distance from the cluster center and the average distance of that cluster. The far away an instance o from the center of it's cluster, the more likely that o is an anomaly instance. Next, we calculate the minimum anomaly score threshold and maximum anomaly score threshold for each cluster using Eq. 3 [36] and Eq. 4 [36], respectively, where $Q1$ represents 25th percentile of the data, $Q3$ represents 75th percentile of the data. The interquartile range (IQR) is calculated by the difference between $Q3$ and $Q1$ as shown in Eq. 5 [36]. Finally, an instance is detected as an anomaly having an anomaly score greater than max threshold (MaxT) or less than min threshold (MinT):

$$\text{Min threshold (MinT)} = Q1 - 1.5 \times \text{IQR} \tag{3}$$

$$\text{Max threshold (MaxT)} = Q3 + 1.5 \times \text{IQR} \tag{4}$$

$$\text{Interquartile range (IQR)} = Q3 - Q1. \tag{5}$$

Algorithm 1 K-means clustering based algorithm for detecting anomalies in heart disease data

```
Input: $D$, a dataset containing $n$ instances
Output: A list of Anomaly or Normal
Method:
for a range of $K$ value do
    Calculate Silhouette score;
End for
Plot out a $K$ value vs silhouette graph;
Find out the optimal value of $K$, which is the highest value of the graph;
Randomly choose $K$ instance from $D$ as the initial cluster centers;
repeat
    Re-assign each instance to the cluster, based on the mean value of the instances in the cluster;
    Update the cluster means;
until no change in cluster means
for each instance of $D$ do
    Calculate anomaly score using Eqn. 2
End for
for each cluster do
    Calculate Min Threshold value ($MinT$) and Max Threshold value($MaxT$) using Eqn. 3 and Eqn. 4 respectively;
End for
for each instance of $D$ do
    if anomaly score > $MaxT$ or anomaly score < $MinT$ then
        The instance is labeled as Anomaly;
    else Normal;
End for
```

Anomaly Removal Module

In our model, once the anomalies are identified, we remove these anomalies from the given dataset to effectively build the prediction model. The reason is that the anomalies or the outliers are typically the data objects that stand out amongst other objects in the dataset and do not conform to the normal behavior in a dataset, which impact on the resultant prediction model. The anomaly detection process for finding such outliers in a given dataset is explained briefly in the earlier section.

Prediction Module

In this module, we apply five popular machine learning classifiers [31] to the heart disease dataset with and without anomalies to test changes in classification results while predicting heart disease. Besides, our proposed anomaly detection model's effectiveness is also evaluated for different cluster values of K.

1. *K-nearest neighbor* (*KNN*) The KNN [1] algorithm is based on the idea that similar things exist close to each other. This closeness is calculated mostly by straight line distance in several existing methods. Among those

approaches, the Euclidean distance method is most popular. At first, the KNN algorithm calculates the distance between the current example of the data point and the query example of the data point. Next, it stores the index and each example's distance to a collection in sorted order. Finally, it classifies all the instances as the mode of the first K labels from the sorted collection. KNN is a pretty intuitive and straightforward non-parametric algorithm. It has no training step, and it continually evolves. On the other hand, KNN can be slow for humongous data and struggle to predict when the dataset's dimensionality is very high. Besides, selecting optimal numbers of K for KNN is one of the most significant issues.

2. *Random forest* (*RF*) The random forest [3] is a supervised classification algorithm based on the decision tree model. Using a bootstrap sampling approach, it first creates K different training data subsets from the original dataset. After that, it builds K decision trees by training these subsets. Finally, a random forest is generated using those decision trees. All the decision trees predict the classification of each sample of the testing dataset depending on the votes of these trees. Random forest uses the ensemble learning technique to reduce the overfitting problem in decision trees and reduces the variance and thus improves the accuracy. On the contrary, it creates many trees, which requires much more computational power and resources and requires much more training time.
3. *Support vector machine* (*SVM*) Support vector machine [5] is a supervised classification algorithm that can be used for both classification or regression challenges. At first, the SVM classifier plots each data point in n-dimensional space where n is the number of features of a dataset. The value of each feature works as the value of a particular coordinate. Finally, it finds out the hyper-plane that differentiates the two classes very well and performs classification. SVM prevents overfitting problems using a good generalization technique and can efficiently handle non-linear data using the Kernel trick. On the other hand, SVM requires longer training time for a larger dataset, and choosing an appropriate Kernel function is difficult.
4. *Naive Bayes* (*NB*) Naive Bayes is a classification algorithm based on Bayes' probability theorem. At first, it converts a dataset into a frequency table. Next, it creates a likelihood table by finding the probabilities for each category of a feature. Then, it uses Bayes' formula to calculate the posterior probability for each class. Finally, the class having the highest posterior probability is the outcome of the prediction. Naive Bayes requires shorter training time since it can estimate test data from a small amount of training data. On the other hand, it implicitly assumes that all the predictors are mutually independent. However, it is impossible to find a set of entirely independent predictors.
5. *Logistic regression* (*LR*) Logistic regression [6] is a supervised classification algorithm that is used to model the probability of a class for each test instance using a logistic function such as the Sigmoid function. Sigmoid function, which has the characteristic of an S-shaped curve, maps any real-valued number between the range of 0 and 1 and later transforms those values into either 0 or 1. Logistic regression performs better when the dataset is linearly separable and less prone to overfitting. On the contrary, the linearity assumption between the dependent variable and the independent variables is the principal disadvantage of logistic regression.

Experimental Evaluation

This section defines the performance metrics that are used in this study to evaluate all classification models in terms of heart disease prediction. We use the following evaluation metric like *precision*, *recall*, *accuracy* to show the experimental results:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{8}$$

where true positives (TP) refer to the positive instances that were correctly labeled by the classifier, false positives (FP) refer to the positive instances that were incorrectly labeled by the classifier, true negatives (TN) refer to the negative instances that were correctly labeled by the classifier, and false negatives (FN) refer to the negative instances that were incorrectly labeled by the classifier.

Implementation and Experimental Results

All the experiment is done on Intel Core i5 2.50 GHz Processor with 8 GB RAM. The suggested model is implemented in Python with packages scikit-learn under OS Windows 10.

Estimation of Optimal Clustering Value *K* from Heart Disease Dataset

As we stated earlier, we first process our heart disease data by replacing missing values with imputation. After that, we apply the K-means clustering algorithm for clustering all the

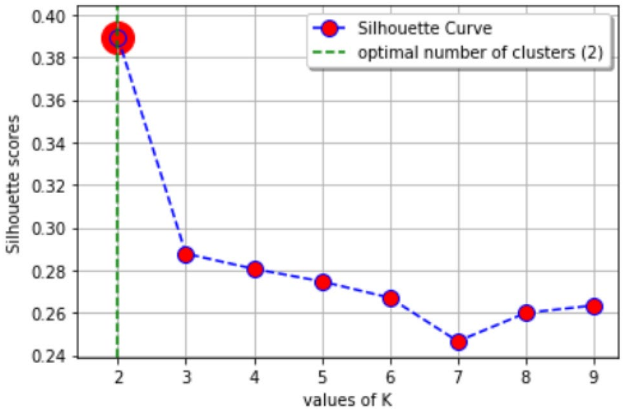

Fig. 3 Plotting the Silhouette scores for different K values

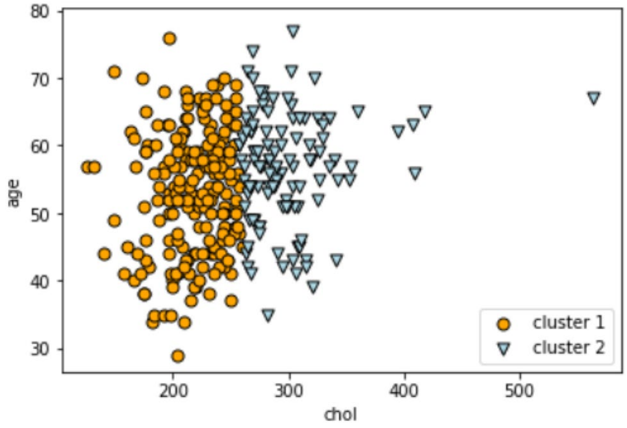

Fig. 4 Scatter plot of "chol" attribute vs. "age" attribute (after clustering as optimal cluster value $K = 2$)

data instances. The K vs. Silhouette score graph was plotted for choosing the optimal K value before applying K-means algorithm. Greater Silhouette score represents the optimal K value for detecting anomalies in heart disease data. From Fig. 3, it is observed that the cluster value K of 2 has the highest Silhouette score.

Apply *K*-Means Clustering Using Optimal Cluster Value of *K* in Heart Disease Data

Next, the K-means clustering algorithm is applied to heart disease data to cluster all data instances. Clustering results from heart disease data using the "chol" attribute vs. "age" attribute is shown in Fig. 4.

Effect of Anomaly Score and Threshold Value

We determine each cluster's mean after clustering all the data instances to measure the anomaly score for each data instance. Next, we measure the max threshold (MaxT) and min threshold (MinT) values using Eqs. 3 and 4 for each cluster. Table 2 shows the threshold values for $K = 2$. Finally, each instance with a score greater than MaxT or less than MinT is detected as an anomaly instance. Figure 5 represents the anomalies labeled by a red-colored square. From Fig. 5, it is shown that our anomaly detection algorithm conducted on heart disease data can separate anomaly instances from normal instances.

Table 2 Max threshold (MaxT) and min threshold (MinT) scores for anomalies in heart disease data using $K = 2$

Cluster no.	MinT	MaxT
One	0.022	1.87
Two	– 0.036	1.81

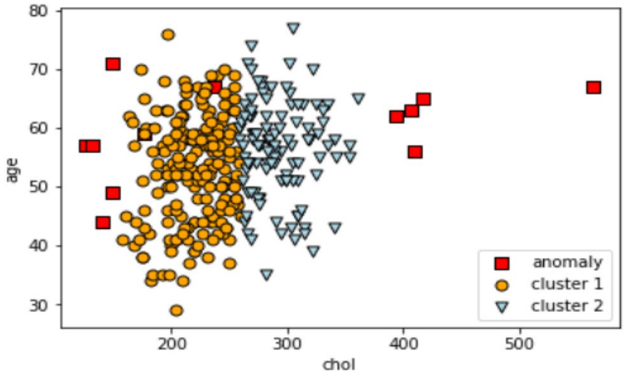

Fig. 5 Scatter plot of "chol" attribute vs. "age" attribute (after anomaly detection)

Testing the Exactness of Detecting Anomalies in Heart Disease Data

After detecting anomalies, anomaly instances are removed by simply looping through all the instances. Then we apply the boxplot method in the heart disease dataset for testing the exactness of anomalies detected by our proposed model. Anomaly detection for the "chol" attribute using boxplot is shown in Fig. 6.

In Fig. 6a, a boxplot of heart disease data with anomalies are represented, and a boxplot of heart disease data without anomalies are represented In Fig. 6b. Figure 6b proves the success of our anomaly detection model for heart disease data.

We also prove our anomaly detection's exactness by scatter plot for "chol" attribute vs. "age" attribute. A scatter plot of original heart disease data with anomalies (labeled as a red-colored square) is shown in Fig. 7. Figure 8 shows the scatter plot of the same dataset after applying our K-means-based anomaly detection model. We can see that our proposed model removes all the anomaly instances successfully.

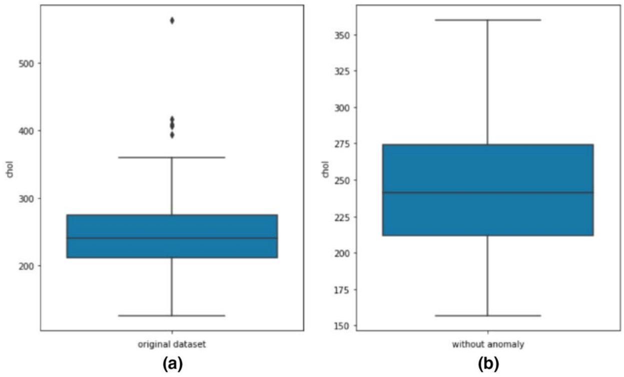

Fig. 6 Boxplot of “chol” attribute

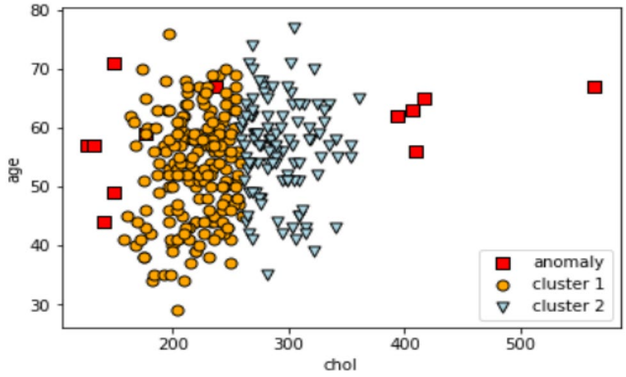

Fig. 7 Scatter plot of “chol” attribute vs. “age” attribute (original dataset)

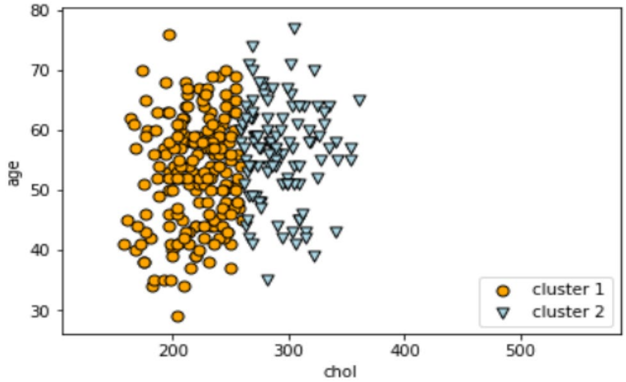

Fig. 8 Scatter plot of “chol” attribute vs. “age” attribute (without anomaly)

Effectiveness Analysis of Our Proposed Anomaly Detection Model Using Optimal *K*

Table 3 Comparison of precision, recall, accuracy

Dataset	Classification models	Accuracy	Precision	Recall
Original dataset	KNN	0.69	0.69	0.69
	RF	0.82	0.83	0.82
	SVM	0.76	0.76	0.76
	NB	0.84	0.84	0.84
	LR	0.81	0.81	0.81
Without anomaly	KNN	0.69	0.68	0.69
	RF	0.88	0.87	0.87
	SVM	0.81	0.80	0.79
	NB	0.84	0.85	0.82
	LR	0.85	0.86	0.84

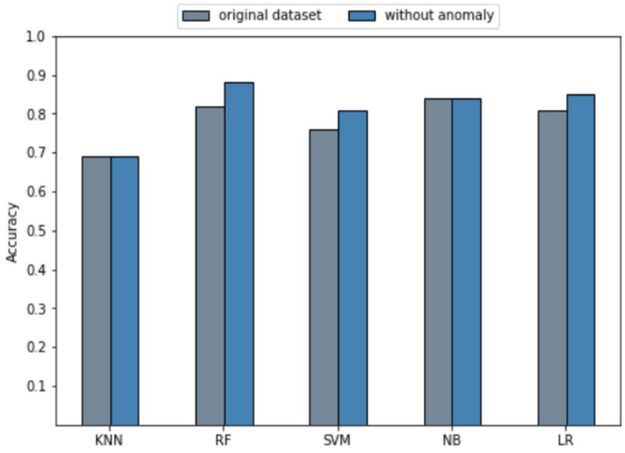

Fig. 9 Comparison of classification accuracy among different classification algorithms

We apply five classification models on heart disease data with and without anomalies to measure our proposed anomaly detection model’s performance. The classification models that are applied are *K*-nearest neighbor (KNN), random forest (RF), support vector machines (SVM), naive-Bayes (NB), logistic regression (LR) to evaluate the proposed models in terms of accuracy, precision, and recall metrics. Table 3 presents the performance comparison of five different classifiers on heart disease data with and without anomaly instances. We see that the performance of RF, SVM, LR classifiers are better in the dataset with no anomaly instance compared with the performance in the original dataset with anomaly instances. The other two classifiers have the same accuracy values for the dataset with and without anomalies. We also observe that RF outperforms other classifiers in terms of accuracy, precision, and recall values for the experiment results on the dataset without anomalies. It proves the effectiveness of our proposed *K*-means-based anomaly detection model for heart disease data (Fig. 9).

Besides, to evaluate the performance of our anomaly detection model, the receiver operating characteristic (ROC) curve of five classifiers for the dataset with and without

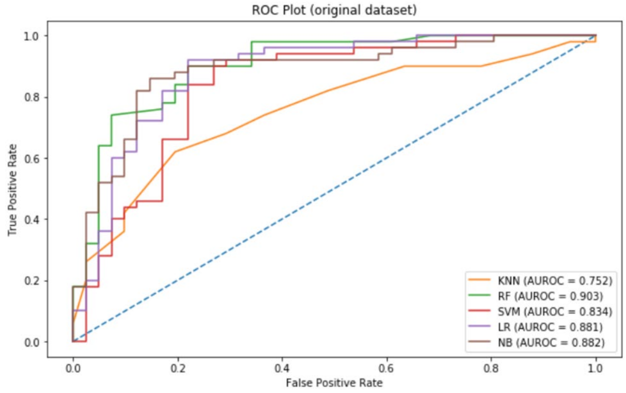

Fig. 10 ROC curve for different classification algorithms (original dataset)

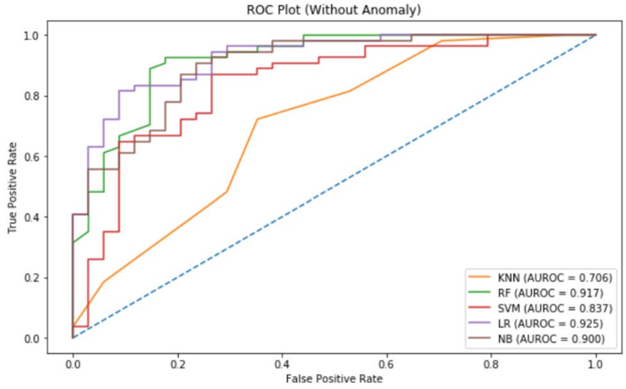

Fig. 11 ROC curve for different classification algorithms (without anomaly)

anomalies have been shown in Figs. 10 and 11. From Fig. 10 it can be observed that AUROC (area under ROC curve) scores for KNN, RF, SVM, LR, NB are 0.752, 0.903, 0.834, 0.881, and 0.882, respectively. After removing anomaly, From Fig. 11, it can be observed that, AUROC scores for KNN, RF, SVM, LR, NB are 0.706, 0.917, 0.837, 0.925 and 0.900 respectively. It is observed that RF, SVM, LR, and NB have a better AUROC score of 0.917, 0.837, 0.925, and 0.900, respectively, after removing anomalies.

Effectiveness Analysis of Our Anomaly Detection Model for Different Cluster Values

The number of cluster values of K has a great impact on the outcomes in clustering-based anomaly detection methods. So, it is essential to find the optimal cluster value. The different number of cluster values affect the anomaly detection in heart disease data. The numbers of anomalies (in percentage) detected by our K-means-based anomaly detection model for different cluster values are shown in Table 4. The relation between the cluster value and the number of anomalies detected by our model is not linear. In our study, we considered anomalies that are greater than or less than some threshold values (MaxT or MinT, respectively) for a particular cluster. As cluster value increases, in most cases, anomalies themselves form their own clusters and count some anomalies as normal data instances. Again, As cluster value increases, clusters can be formed from normal instances also. Then, some normal instances can be detected as anomaly according to the threshold values of that cluster. The comparison of the impact of cluster values in detecting anomalies are shown in Figs. 12 and 13 for $K = 3$ and $K = 4$, respectively. From Fig. 12, it is shown that anomalies (red square) are far from the clusters and have a small effect in detecting anomalies. But Fig. 13, it is shown that anomalies are mixed with the normal data instances, and this phenomenon may hamper detecting anomalies.

Table 4 Percentage of anomaly for different cluster value

Cluster value	Anomaly percentage (%)
$K = 2$	3.96
$K = 3$	3.63
$K = 4$	1.98
$K = 5$	3.63
$K = 6$	3.96
$K = 7$	4.62
$K = 8$	2.64
$K = 9$	2.31

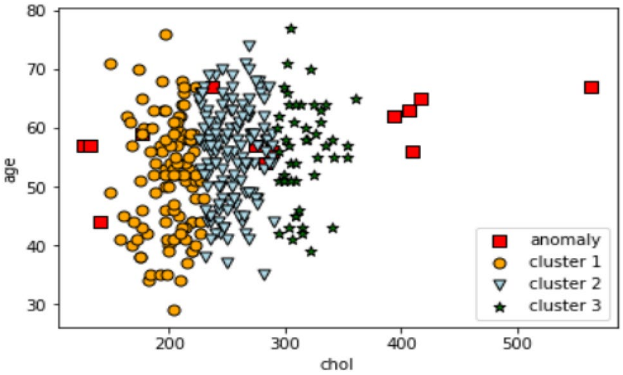

Fig. 12 Scatter plot of “chol” attribute vs. “age” attribute ($K = 3$)

Changes in cluster value can affect the classifications of anomalies in heart disease data. A comparison in terms of accuracy for different cluster values from 2 to 9 has shown in Table 5. From Table 5, it is observed that RF achieve best classification accuracy value of 0.88 for both $K = 2$ and $K = 5$. Besides that, LR achieves the highest accuracy value of 0.88 for $K = 5$ and $K = 8$. In addition, the best accuracy value of 0.88 has been achieved in NB for $K = 8$, and the

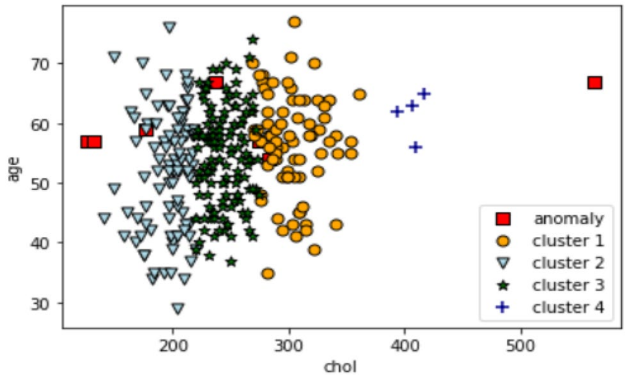

Fig. 13 Scatter plot of “chol” attribute vs. “age” attribute ($K = 4$)

best accuracy value of 0.85 for SVM has been found when $K = 8$. On the other hand, KNN achieved the best accuracy value of 0.77 for $K = 5$. So, from previous analysis, it is observed that our anomaly detection model achieves better performance in most of the classifiers for $K = 8$. As accuracy is never higher than 0.88, our anomaly detection model proves the effectiveness for optimal cluster value $K = 2$ in terms of the trade-off between accuracy and computation cost.

Discussion

Our anomaly detection model is completely based on the K-means clustering algorithm. First, we have estimated optimal cluster value $K = 2$ for heart disease data using the Silhouette method to reduce computational cost. After that, the clustering of heart disease data instances has been done using the K-means algorithm. Then, for every instance of heart disease data, the anomaly score has been calculated using Eq. 2. After that, for each cluster, max threshold (MaxT) and min threshold (MinT) have been calculated using Eqs. 4 and 3, respectively. Finally, instances having a greater anomaly score than MaxT or lesser anomaly score than MinT are detected as an anomaly and removed.

The exactness of our K-means-based anomaly detection model for heart disease data is tested using the boxplot method and scatter plot, in which we have proved our anomaly detection model is successful in finding anomalies as shown in Figs. 6, 7 and 8. Besides, there is an improvement of prediction accuracy for RF, SVM, and LR after removing all the anomaly instances, as shown in Fig. 9. An improvement in the area under the ROC curve (AUROC) is observed for RF, SVM, NB, and LR after removing all the anomaly instances, as shown in Fig. 11.

In addition, the effectiveness analysis of our anomaly detection model for different cluster values has been performed. It is observed that, As cluster value increases, in most cases, anomalies themselves form their own clusters and count some anomalies as normal data instances. Again, As cluster value increases, clusters can be formed from normal instances also. Then some normal instances can be detected as anomaly according to the threshold values of that cluster. Our anomaly detection model achieves better accuracy for $K = 2, 5, 8, 9$. For all cases, the best accuracy is 0.88. But considering the trade-off between accuracy and computation cost, it is obvious that our K-means-based anomaly detection model is effective for optimal cluster value $K = 2$. However, our anomaly detection model's accuracy is better but still sensitive to the trade-off between the number of anomalies detected and classification accuracy. In some cases, it is found that an increase in cluster value K can lead to bad anomaly detection, as shown in Fig. 13. Although we use a heart disease dataset to measure our anomaly detection model's effectiveness, this approach can be applied to other application areas such as IoT analytics [28], cybersecurity [30], etc. Our anomaly detection model could also play a significant role while conducting rule-based analysis based on a given dataset in a particular application domain [25, 29].

Table 5 Comparison of accuracy for different cluster values

Dataset	Cluster value	KNN	RF	SVM	NB	LR
Original dataset		0.69	0.82	0.76	0.84	0.81
Without anomaly	$K = 2$	0.69	0.88	0.81	0.84	0.85
	$K = 3$	0.73	0.86	0.83	0.83	0.84
	$K = 4$	0.73	0.78	0.74	0.80	0.81
	$K = 5$	0.77	0.88	0.84	0.84	0.88
	$K = 6$	0.74	0.77	0.76	0.81	0.78
	$K = 7$	0.76	0.84	0.83	0.80	0.86
	$K = 8$	0.74	0.85	0.85	0.88	0.88
	$K = 9$	0.71	0.82	0.78	0.85	0.83

Conclusion and Future Work

This paper has presented a data-driven prediction model considering the associated risk factors in heart disease, which can play a significant role in the healthcare domain. While building such an effective model based on machine learning techniques, we have considered the quality of the data getting by removing the anomalies from the data in our model. A clustering-based anomalies detection model using optimal clustering value has been used for heart disease data. The effectiveness of our model has been evaluated with and without anomalies using various classifiers. From them, RF, SVM, LR classifiers using without anomalies achieves better accuracy than with anomalies data. Our anomaly detection model recognizes anomalies effectively for different cluster values and achieves better accuracy for different cluster values, and thus it proves the effectiveness of our optimal K-means-based anomaly detection model for building the prediction model. In the future, we will focus on collecting recent datasets from a medical center to do further experiments and analysis to assist the healthcare community.

Compliance with Ethical Standards

Conflict of interest The authors declare no conflict of interest.

References

1. Altman N. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992;46(3):175–85.
2. Ayon SI, Islam MM, Hossain MR. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE J Res. 2020;1–20.
3. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
4. Campello RJ, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: Pacific-Asia conference on knowledge discovery and data mining. Springer; 2013. p. 160–72.
5. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
6. Cramer JS. The origins of logistic regression; 2002.
7. Dessai ISF. Intelligent heart disease prediction system using probabilistic neural network. Int J Adv Comput Theory Eng. 2013;2(3):2319–526.
8. Ding Z, Fei M. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. IFAC Proc Vol. 2013;46(20):12–7.
9. Fan J, Zhang Q, Zhu J, Zhang M, Yang Z, Cao H. Robust deep auto-encoding gaussian process regression for unsupervised anomaly detection. Neurocomputing. 2020;376:180–90.
10. Forgy EW. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics. 1965;21:768–9.
11. Fujimaki R, Yairi T, Machida K. An approach to spacecraft anomaly detection problem using kernel feature space. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining; 2005. p. 401–10.
12. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Amsterdam: Elsevier; 2011.
13. Janakiram D, Reddy V, Kumar AP. Outlier detection in wireless sensor networks using Bayesian belief networks. In: 2006 1st International conference on communication systems software & middleware. IEEE; 2006. p. 1–6.
14. Kumar V. Parallel and distributed computing for cybersecurity. IEEE Distrib Syst Online. 2005;6(10).
15. Liu FT, Ting KM, Zhou ZH. Isolation forest. In: 2008 Eighth IEEE international conference on data mining. IEEE; 2008. pp. 413–22.
16. Mascaro S, Nicholso AE, Korb KB. Anomaly detection in vessel tracks using Bayesian networks. Int J Approx Reason. 2014;55(1):84–98.
17. Mohamed MS, Kavitha T. Outlier detection using support vector machine in wireless sensor network real time data. Int J Soft Comput Eng. 2011;1(2).
18. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access. 2019;7:81542–54.
19. Münz G, Li S, Carle G. Traffic anomaly detection using k-means clustering. In: GI/ITG workshop MMBnet; 2007. p. 13–4.
20. Nachman B, Shih D. Anomaly detection with density estimation. Phys Rev D. 2020;101(7):075042.
21. Ranjith R, Athanesious JJ, Vaidehi V. Anomaly detection using dbscan clustering technique for traffic video surveillance. In: 2015 Seventh international conference on advanced computing (ICoAC). IEEE; 2015. p. 1–6.
22. Ripan RC, Sarker IH, Furhad MH, Anwar MM, Hoque MM. An effective heart disease prediction model based on machine learning techniques; 2020.
23. Ronit: Heart disease uci; 2018. https://www.kaggle.com/ronitf/heart-disease-uci.
24. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
25. Sarker IH. Context-aware rule learning from smartphone data: survey, challenges and future directions. J Big Data. 2019;6(1):95.
26. Sarker IH. A machine learning based robust prediction model for real-life mobile phone data. Internet Things. 2019;5:180–93.
27. Sarker IH, Abushark YB, Alsolami F, Khan AI. Intrudtree: a machine learning based cyber security intrusion detection model. Symmetry. 2020;12(5):754.
28. Sarker IH, Hoque MM, Uddin MK, Alsanoosy T. Mobile data science and intelligent apps: concepts, ai-based modeling and research directions. Mob Netw Appl. 2020;1–19.
29. Sarker IH, Kayes A. Abc-ruleminer: user behavioral rule-based machine learning method for context-aware intelligent services. J Netw Comput Appl. 2020;102762.
30. Sarker IH, Kayes A, Badsha S, Alqahtani H, Watters P, Ng A. Cybersecurity data science: an overview from machine learning perspective. J Big Data. 2020;7(1):1–29.
31. Sarker IH, Kayes A, Watters P. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. J Big Data. 2019;6(1):57.
32. Spence C, Parra L, Sajda P. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In: Proceedings IEEE workshop on mathematical methods in biomedical image analysis (MMBIA 2001). IEEE; 2001. p. 3–10.
33. Sun L, Versteeg S, Boztas S, Rao A. Detecting anomalous user behavior using an extended isolation forest algorithm: an enterprise case study. 2016. arXiv preprint. arXiv:1609.06676.

34. Tax DM, Duin RP. Support vector data description. Mach Learn. 2004;54(1):45–66.
35. Tu B, Yang X, Li N, Zhou C, He D. Hyperspectral anomaly detection via density peak clustering. Pattern Recognit Lett. 2020;129:144–9.
36. Wickham H, Stryjewski L. 40 years of boxplots. Am. Stat. 2011.
37. Xu J, Shelton CR. Intrusion detection using continuous time Bayesian networks. J Artif Intell Res. 2010;39:745–74.
38. Xue Z, Shang Y, Feng A. Semi-supervised outlier detection based on fuzzy rough c-means clustering. Math Comput Simul. 2010;80(9):1911–21.
39. Yoon KA, Kwon OS, Bae DH. An approach to outlier detection of software measurement data using the k-means clustering method. In: First international symposium on empirical software engineering and measurement (ESEM 2007. IEEE; 2007. p. 443–5.
40. Zhang C, Song D, Chen Y, Feng X, Lumezanu C, Cheng W, Ni J, Zong B, Chen H, Chawla NV. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33; 2019. p. 1409–16.