 doi:10.20944/preprints202011.0744.v1

An Effective Heart Disease Prediction Model based on Machine Learning Techniques

Rony Chowdhury Ripan[1], Iqbal H. Sarker[1,*], Md. Hasan Furhad[2], Md Musfique Anwar[3], and Mohammed Moshiul Hoque[1]

[1] Dept of Computer Science & Engineering,
Chittagong University of Engineering & Technology,
Chittagong-4349, Bangladesh.
[2] Canberra Institute of Technology,
Canberra, Australia.
[3] Jahangirnagar University, Dhaka, Bangladesh.
∗Correspondence: iqbal@cuet.ac.bd

Abstract. This paper presents an effective heart disease prediction model through detecting the anomalies, also known as outliers, in healthcare data using the unsupervised *K-means clustering* algorithm. Most existing approaches for detecting anomalies are based on constructing profiles of normal instances. However, such techniques require an adequate number of normal profiles to justify those models. Our proposed model first evaluates an *optimal* value of K using Silhouette method. Next, it intends to locate anomalies that are far from a certain threshold distance with respect to their clusters. Finally, the five most popular classification techniques such as K-Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machines (SVM), Naive Bayes (NB), and Logistic Regression (LR) are applied to build the resultant prediction model. The effectiveness of the proposed methodology is justified using a benchmark dataset of heart disease.

Keywords: Anomaly detection, Healthcare, K-means clustering, Heart disease prediction, Data analytics, Machine learning

1 Introduction

The modern digital world is overwhelmed by an unprecedented amount of data in various domains [17]. Most organizations usually have no problem in capturing the ample amount of data, but the challenging task for them is to elucidate and extract the required meaningful knowledge or information from these vast amount of data in an efficient manner. Several data mining and machine learning techniques are used to find interesting and meaningful relationships among data [14] [18]. One such technique is known as anomaly detection. Anomalies are the data characteristics that are different from normal behaviors [19]. In some cases, such anomalies are considered as noise or outliers that affects on a machine learning based prediction model [15]. Detection of anomalies has recently occupied

 doi:10.20944/preprints202011.0744.v1

an overwhelming research interest owing to its necessity in various domains to get critical actionable information from large datasets.

Recently, some studies have focused on anomaly detection from larger datasets [10, 21, 3, 16]. An elaborate discussion on these methodologies is given in Section 2. The common phenomenon of most of the existing research works is to construct a profile of normal instances which is considered as challenging task as it requires to find a sufficient number of normal profiles.

In this study, we present a model to *detect anomalies* in the healthcare domain based on the *K-means clustering* algorithm where the optimal value of K is measured using Silhouette method. The major advantage of our proposed approach is that it does not require to construct the normal profiles or knowledge of previous anomaly records in the heart disease training dataset. Usually, the anomalous instances lie in sparse or small clusters and thus far from the centroids of their respective clusters. We apply five popular machine learning classification techniques [20] such as K-Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machines (SVM), Naive-Bayes (NB), and Logistic Regression (LR) to measure the effectiveness of the proposed methodology in anomaly detection in order to predict heart disease.

The rest of this paper is organized as follows: we present the related works on anomaly detection and heart disease prediction in Section 2. Section 3 provides the details of the proposed anomaly detection model. Our experimental results with discussions are covered in Section 4. Finally, we conclude the paper in Section 5 by providing the future directions of this research.

2 Related Work

A considerable amount of research has been devoted to the problem of anomaly detection. For instance, Janakiram et al. [6] proposed a Bayesian belief networks based anomaly detection to detect the anomalies. Steven Mascaro et al. [7] proposed a model that detects anomalies using Bayesian networks. Their Bayesian network model learns anomalies from real-world Automated Identification System (AIS) data.

Zhiguo et al. [2] presented a novel anomaly detection framework based on Isolation Forest, in which they used the frame of sliding windows and also considered the concept of drift phenomenon. Ranjith et al. [11] proposed an unsupervised anomaly detection model using the DBSCAN algorithm. They tried to find out anomalies from a traffic dataset, in which a trajectory is said to be an anomaly if it does not fit with the trained model. Munz et al. [9] also tried to find anomalies in a traffic dataset using K-means clustering algorithm.

There are numerous studies on heart disease detection. Safial Islam et al. [1] did a comparative study on coronary artery heart disease prediction, in which they applied several data mining techniques like 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. [8] proposed a novel method that aims to improve the prediction

accuracy by finding significant features. The prediction model considered different combinations of features and then the performance of the model is evaluated by applying several known classification techniques.

Our proposed anomaly detection model is based on unsupervised approach where we apply the optimal K-means clustering algorithm to cluster anomalies in the heart disease data. The optimal cluster value of K has been estimated using Silhouette method, and classification techniques are used to effectively predict heart disease by removing anomalies.

3 Methodology

In this section, we present the details of our proposed anomaly detection model, which has four different modules as presented in Fig. 1.

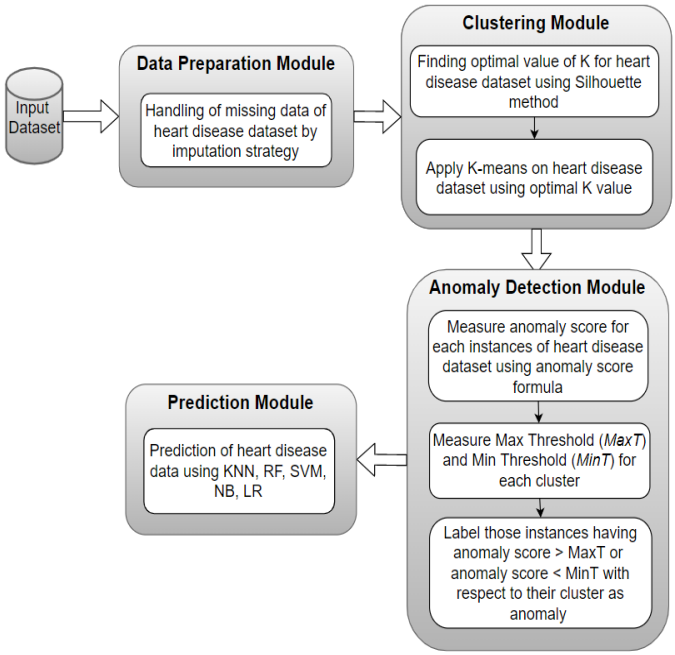


Fig. 1: Proposed model for effectively predicting heart disease.

3.1 Data Preparation Module

Instances in healthcare datasets typically contain several healthcare features and related facts that can be used to build the model. In this work, we use a

benchmark heart disease dataset available in Kaggle [12]. This dataset contains 303 instances. Each instance has 13 features of which twelve are "Integer" type and one is "Float" type.

Our heart disease dataset has a small number of attributes with missing values and we use the imputation strategy. There are several ways to impute missing values. In our approach, we impute the mean value for each attribute.

3.2 Clustering Module

K-means Clustering: The K-means algorithm [4] is an unsupervised clustering algorithm. It takes the number of clusters and the dataset as input and gives the output as a set of clusters. The K-means algorithm determines the center of a cluster as the mean value of the instances within that cluster. First, it randomly selects K from the instances, which represents a center or cluster means in the dataset. Each of the residual instances are assigned to the nearest cluster based on the Euclidean distance between the instance and the cluster mean. Next, it recurrently optimizes the positions of the centers for all the clusters.

For each cluster, it calculates the new mean as a center using the instances assigned to that cluster in the previous iteration. All the instances of each cluster are then reassigned to the updated means. The iterations continue until the algorithm has converged, i.e., centers of all the clusters don't need any more repositioning. Different values of K can lead to different results, and so it is important to find an optimal value of K.

In this study, we apply the Silhouette method to find the optimal value of K from the heart disease dataset. Given a range of K values, the Silhouette method computes the Silhouette score, i.e., *Silhouette Coefficient* for all the instances. The Silhouette Coefficient for an instance is calculated using Eqn. 1 [13]. In this equation, a indicates the mean distance between the instances within-cluster, and b is the mean distance between the instance and the nearest cluster/s. The Silhouette Coefficient value ranges from -1 to +1, where +1 indicates the best cluster fit and -1 means the worst cluster fit.

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

3.3 Anomaly Detection Module

In our proposed approach, we calculate the anomaly score of an instance (as shown in Eqn. 2) based on the distance between the instance and the center of its nearest cluster [5].

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

In this formula, *distance(o,C_o)* represents the distance between instance o and cluster center C_o, whereas L indicates the mean distance of that cluster. So *Anomaly Score* in Eqn. 2 measures the ratio of the distance of each instance

from the cluster center to the mean distance of that cluster. The further 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 threshold score and maximum anomaly threshold score for each cluster using Eqn. 3 [22] and Eqn. 4 [22], respectively, where $Q1$ represents $25th$ percentile of the data and $Q3$ represents $75th$ percentile of the data. The Interquartile range (IQR) is the difference between $Q3$ and $Q1$ as shown in Eqn. 5 [22].

Finally, all the instances having an anomaly score greater than Max Threshold ($MaxT$) or less than Min Threshold($MinT$) will be detected as an anomaly.

$$Min\ Threshold(MinT) = Q1 - 1.5 * IQR \tag{3}$$

$$Max\ Threshold(MaxT) = Q3 + 1.5 * IQR \tag{4}$$

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

3.4 Prediction Module

In this module, we apply five different classifiers [20] such as Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and K-Nearest Neighbor (KNN) on the heart disease dataset with and without anomalies to measure the effectiveness of our model.

4 Experimental Evaluation

4.1 Experimental Setup

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

4.2 Implementation and Experimental Results

As stated earlier, we first process our heart disease data by replacing missing values with imputation.

Then, we plot the K vs. Silhouette score graph (as shown in Fig. 2) in order to choose the optimal K value before applying the K-means algorithm. It is observed from Fig. 2 that the cluster value K of 2 has the highest Silhouette score.

Next, the K-means clustering algorithm is applied to heart disease data to cluster all data instances. After clustering all the data instances, we determine the mean of each cluster in order to measure the anomaly score of each data instance. Next, we measure the Max Threshold ($MaxT$) and Min Threshold ($MinT$) values using Eqn. 3 and Eqn. 4 for each cluster. For cluster 1, it is observed that $MinT$ and $MaxT$ are 0.022 and 1.87, respectively. For cluster 2, it is observed that $MinT$ and $MaxT$ are -0.036 and 1.81, respectively. Finally,

 doi:10.20944/preprints202011.0744.v1

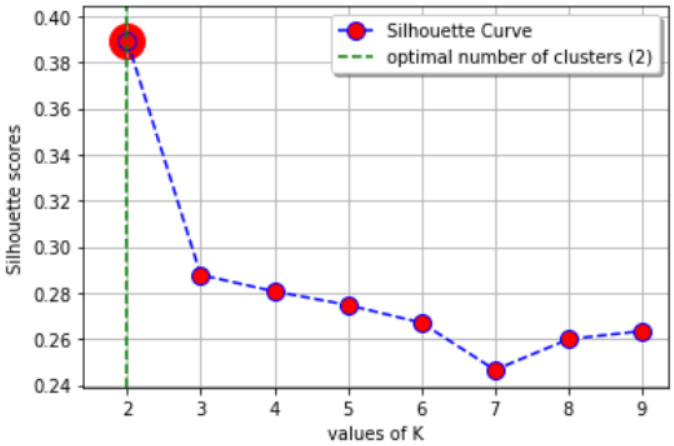


Fig. 2: Plotting the Silhouette scores for different K values

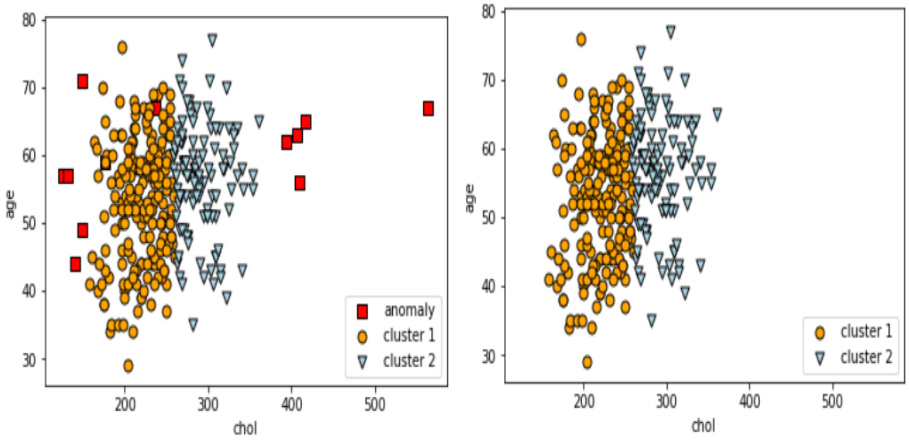
each instance that has a score greater than $MaxT$ or less than $MinT$ is detected as an anomaly instance and therefore removed those instances. A scatter plot of original heart disease data with anomalies (labeled as a red-colored square) is shown in Fig. 3a. Fig. 3b shows that our proposed model removes all the anomaly instances successfully.

We apply five classification models on heart disease data with and without anomalies to measure the performance of our proposed anomaly detection model. The classification models that are applied are K-Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machines (SVM), Naive-Bayes (NB) and Logistic Regression (LR) to evaluate the proposed models in terms of accuracy, precision, and recall metrics.

Table 1 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 achieve similar accuracy values for the dataset with and without anomalies. In Fig. 4, We also observe that RF outperforms other classifiers in terms of accuracy, precision, and recall values for the experiment results performed on the dataset without anomalies. Thus, experiment results prove the effectiveness of our proposed K-means based anomaly detection model for heart disease data.

In addition, we plot the Receiver Operating Characteristic (ROC) curve of five classifiers for the dataset with and without anomalies to evaluate the performance of our anomaly detection model as shown in Fig. 5a and in Fig. 5b. From 5b, it is proven that RF, SVM, LR, and NB have a better area under the ROC curve (AUC) of values 0.917, 0.837, 0.925, and 0.900, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 November 2020 doi:10.20944/preprints202011.0744.v1



(a) Original Dataset (b) Without anomaly

Fig. 3: Scatter plot of Heart Disease Data

Table 1: Comparison of Precision, Recall, and 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

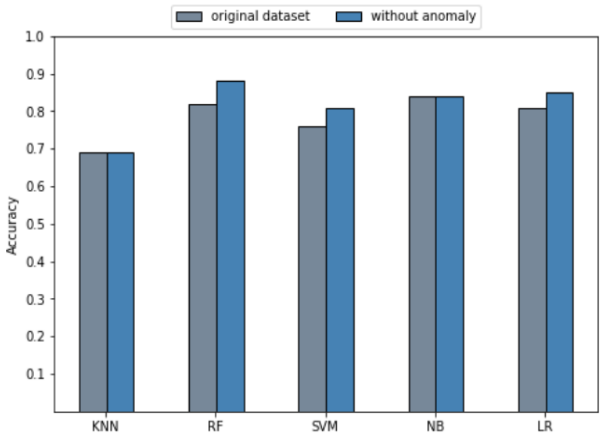
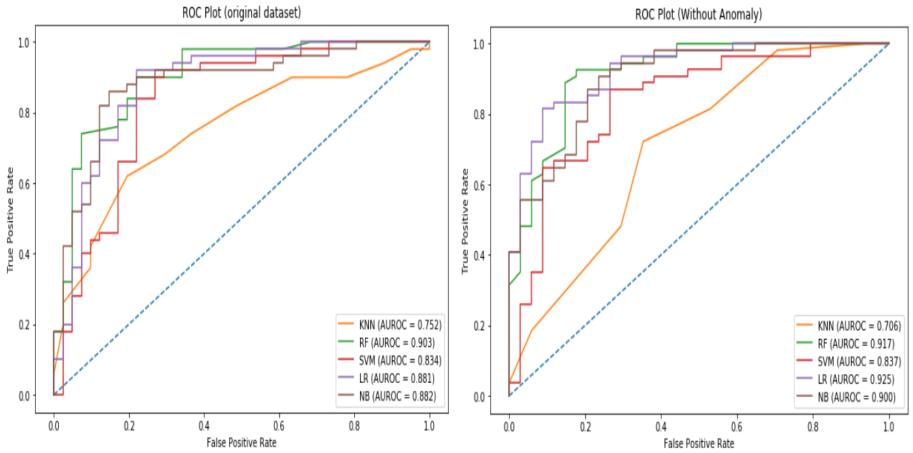


Fig. 4: Comparison of classification accuracy among different classification algorithms



(a) original dataset (b) without anomaly

Fig. 5: ROC curve with various machine learning models

5 Conclusion

We have presented an effective heart disease prediction model based on machine learning techniques. The effectiveness of our model has been evaluated with and without anomalies using various classifiers. Experiment results showed that RF, SVM, LR classifiers achieved better accuracy in the dataset without anomalies compared with dataset with anomaly instances. Again, our anomaly detection model is able to effectively recognize the anomalies in the data. In future, we will focus on additional experiments to measure the effectiveness of our model, and also on the model effectiveness in other application areas like IoT systems.

References

1. Safial Islam Ayon, Md Milon Islam, and Md Rahat Hossain. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. *IETE Journal of Research*, pages 1–20, 2020.
2. Zhiguo Ding and Minrui Fei. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. *IFAC Proceedings Volumes*, 46(20):12–17, 2013.
3. Jinan Fan, Qianru Zhang, Jialei Zhu, Meng Zhang, Zhou Yang, and Hanxiang Cao. Robust deep auto-encoding gaussian process regression for unsupervised anomaly detection. *Neurocomputing*, 376:180–190, 2020.
4. Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.
5. Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
6. Dharanipragada Janakiram, VA Reddy, and AVU Phani Kumar. Outlier detection in wireless sensor networks using bayesian belief networks. In *2006 1st International Conference on Communication Systems Software & Middleware*, pages 1–6. IEEE, 2006.
7. Steven Mascaro, Ann E Nicholso, and Kevin B Korb. Anomaly detection in vessel tracks using bayesian networks. *International Journal of Approximate Reasoning*, 55(1):84–98, 2014.
8. Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7:81542–81554, 2019.
9. Gerhard Münz, Sa Li, and Georg Carle. Traffic anomaly detection using k-means clustering. In *GI/ITG Workshop MMBnet*, pages 13–14, 2007.
10. Benjamin Nachman and David Shih. Anomaly detection with density estimation. *Physical Review D*, 101(7):075042, 2020.
11. R Ranjith, J Joshan Athanesious, and V Vaidehi. Anomaly detection using dbscan clustering technique for traffic video surveillance. In *2015 Seventh International Conference on Advanced Computing (ICoAC)*, pages 1–6. IEEE, 2015.
12. Ronit. Heart disease uci, Jun 2018.
13. Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
14. Iqbal H Sarker. Context-aware rule learning from smartphone data: survey, challenges and future directions. *Journal of Big Data*, 6(1):95, 2019.

15. Iqbal H Sarker. A machine learning based robust prediction model for real-life mobile phone data. *Internet of Things*, 5:180–193, 2019.
16. Iqbal H Sarker, Yoosef B Abushark, Fawaz Alsolami, and Asif Irshad Khan. Intrudtree: A machine learning based cyber security intrusion detection model. *Symmetry*, 12(5):754, 2020.
17. Iqbal H Sarker, Mohammed Moshiul Hoque, Md Kafil Uddin, and Tawfeeq Alsanoosy. Mobile data science and intelligent apps: Concepts, ai-based modeling and research directions. *Mobile Networks and Applications*, pages 1–19, 2020.
18. Iqbal H Sarker and ASM Kayes. Abc-ruleminer: User behavioral rule-based machine learning method for context-aware intelligent services. *Journal of Network and Computer Applications*, page 102762, 2020.
19. Iqbal H Sarker, ASM Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters, and Alex Ng. Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*, 7(1):1–29, 2020.
20. Iqbal H Sarker, ASM Kayes, and Paul Watters. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data*, 6(1):57, 2019.
21. Bing Tu, Xianchang Yang, Nanying Li, Chengle Zhou, and Danbing He. Hyperspectral anomaly detection via density peak clustering. *Pattern Recognition Letters*, 129:144–149, 2020.
22. Hadley Wickham and Lisa Stryjewski. 40 years of boxplots. *Am. Statistician*, 2011.