Toward a High-Accuracy Hybrid System for Cardiac Patient Data Analysis using C-Means Fuzzy Clustering in Neural Network Structure

Mahmoud Karim Qaseem[1], Razieh Asgarnezhad[2]
1- Department of Computer Engineering, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran
Email: Mahmoudalmutairi89@gmail.com
2- Department of Computer Engineering, Aghigh Institute of Higher Education Shahinshahr, 8314678755, Isfahan, Iran
Email: razyehan@gmail.com (Corresponding author)

Received: 11 January 2023 Revised: 19 February 2023 Accepted: 28 March 2023

ABSTRACT:
The main problem related to heart disease is the lack of timely diagnosis or the general weakness in the diagnosis of this disease, which is also due to the lack of selection of the appropriate model by the doctor or the lack of proper use of standard models. One of the essential applications of data mining techniques is related to medicine and disease diagnosis. One of the data mining techniques is information clustering. This paper will try to provide a model for the diagnosis of heart disease and its improvement in terms of accuracy on the standard UCI heart database. In this research, with a comprehensive and complete review of the C-Meaning fuzzy clustering method and neural networks in the field of heart disease prediction, an attempt is made to improve these solutions and provide new solutions in this field. The main goal is to combine these two data mining algorithms, both of which alone showed the highest accuracy and the fastest speed in past research. The current authors are trying to find a model that has higher accuracy and speed than the previous methods and makes fewer mistakes and has significantly higher efficiency than other models. The numerical tests implemented on the proposed model show the superiority of the new model compared to the conventional methods in the literature.

KEYWORDS: Data Analysis, Cardiac Patients, Hybrid System, High Accuracy, C-Means Fuzzy Clustering, Neural Network Structure

1. INTRODUCTION

The learning problem is to find a general rule to describe the data using a limited number of collected samples. Nowadays, learning methods are used in many fields such as data mining (DM), computer science, statistics, etc., and many researches are being conducted on them. The new knowledge of DM is one of the developing knowledge that has established its position in all fields in recent years, so that its growth is increasing compared to other superior knowledge.

Health and treatment is one of the areas that has been the focus of DM experts in recent years. One of the most important challenges of medical organizations is to determine the quality of their services in front of affordable cost for the users of these services. The quality of service depends on the correct diagnosis and prescription of correct behavior, and poor clinical decisions can lead to adverse outcomes that are unpredictable. The quality of services includes the correct identification of diseases and their effective treatment. Poor clinical decisions can lead to dire consequences that are unacceptable. Also, hospitals should minimize the cost of clinical tests. This can be achieved by using computer-based information and decision-making systems. Today, many hospitals use some types of Hospital Information System to manage patient data or healthcare [1].

These systems generate a large amount of data including numbers, texts, diagrams, photos. Unfortunately, these data are rarely used in clinical decision-making. There is a valuable capital of hidden information in this data. The question that has occupied the minds of most decision makers in this field is how we can turn data into useful information that enables the medical staff to make intelligent clinical decisions.

An important group of problems in medicine is related to the diagnosis of diseases, which is performed

Paper type: Research paper
DOI: 10.30486/MJTD.2023.1977501.1014
How to cite this paper: M. Karim Qaseem, R. Asgarnezhad, "**Toward a High-Accuracy Hybrid System for Cardiac Patient Data Analysis using C-Means Fuzzy Clustering in Neural Network Structure**", *Majlesi Journal of Telecommunication Devices*, Vol. 12, No. 2, pp.53-67, 2023.

on the basis of various tests on the patient. When the number of parameters in disease diagnosis increases, it may be difficult to diagnose the disease even for an expert medical expert, which is the reason why computer diagnostic tools have been used for the purpose of helping medicine in the last few decades. By discovering hidden patterns among this information, the decision-making process can be improved. The knowledge extracted from these data can be provided to doctors as a decision support system (DSS).

Medical data is very diverse and extensive, including patient data, resource management data, costs, etc. In order to achieve success, medical organizations must have the ability to analyze data properly. On the other hand, medical environments are basically rich environments in terms of information. In the meantime, the field of heart diseases is of double importance due to the sensitivity of heart health in the continuation of human life, and improving diagnoses and treatments in this field can save many lives. In this study, the DSS in the field of heart disease diagnosis and the suggestion of treatment methods for patients is designed based on data mining.

A heart attack (in the medical term MI) or Myocardial infarction or heart attack is permanent and irreversible destruction and cell death in a part of the heart muscle (myocardium) due to the loss of blood flow and the occurrence of a severe ischemia in that part of the heart. This stoppage of blood circulation may appear suddenly without any previous symptoms or appear after several angina attacks (chest pain). The main cause of stroke is the closing of the blood vessels feeding the heart. Balloons and open heart surgery (blocked vessel replacement) are used to remove blockages other than drugs. Heart attack is a widespread complication that causes thousands of deaths every year [2].

The main DM techniques are divided into descriptive and predictive categories. In the case of heart disease diagnosis, DM techniques with the purpose of prediction are discussed. Therefore, data analysis methods have become a useful assistant for doctors to make heart disease diagnosis decisions. Research results show that DM technology has been successfully implemented in the prediction of heart disease. Various techniques have been proposed to predict heart disease. The number of features increases the computational time and decreases the accuracy. Therefore, a set of optimal features should be obtained. Due to the computing time, feature filtering and classification with high accuracy has become one of the new topics. Various methods have been used for classification such as decision trees, multiple linear regression, logistic regression, audit analysis, simple neural networks (NN), support vector machine (SVM), k-nearest neighbor (KNN), etc. There are different classification methods and algorithms. These methods differ from each other in the architecture, learning algorithm, or the way of training and displaying features, although some classifiers perform better than others, but no classifier can always be superior to the other in any situation, or Classify the data without any errors.

Various researches have been done using clustering algorithms. Algorithms such as K_Means (KM), C-Means Fuzzy (CMF) and KM algorithm is a simple iterative method, and it is used for clustering a set of data in a specific number of clusters (K) that the user determines. KM algorithm has been expressed by many researchers in different ways.

In addition to this, another method that has been used is the method of combining different DM techniques. For example, the combination of genetic algorithm and neural network. The results obtained from the research show that the accuracy and efficiency of the combined methods are far higher than the use of a single algorithm, therefore, in this research, by examining the various methods proposed in the field of data mining using neural networks based on clustering CMF has created a model based on a heart data set so that heart disease can be predicted with the help of this method.

The main goal of this research is to focus on the algorithms and methods presented in data mining for early and accurate diagnosis of heart disease and to try to improve these methods. The main idea considered in this research is the integrated use of neural networks based on CMF clustering to improve performance, which will be evaluated by implementing and testing the performance of the proposed algorithm.

The innovation of this research are of concern:

- Conducting a comprehensive research in the field of activities for quick and accurate prediction of heart disease
- Presenting a new solution based on the use of neural networks based on CMF clustering for quick and accurate prediction of heart disease.
- Presenting a method for quick and accurate prediction of heart disease using neural networks based on CMF clustering.

The rest of this article is of concern: A review of the use of DM in health and treatment has been done with focusing on the heart disease in Sect. 2. Then, topics related to DM used in this research, are presented in Sect. 3. After it, the research method will be explained in Sect. 4. DM analysis will be done using effective tools and models in Sect. 5. Finally, conclusions and future suggestions are presented in Sect. 6.

2. RELATED WORK

Simple Bayes algorithm and KM algorithm is one of the best methods to diagnose heart disease. Various researches have been done in this field. In many researches, these algorithms have been used alone or compared with another algorithm.

In [3], the heart disease analysis system has been

investigated using DM techniques. In order to cluster the pre-processed data in the database, clustering algorithms such as K-means have been used. Also, the Mafia algorithm (MAFIA) has been used to explore the maximum repeating patterns in the heart disease database. The C4.5 algorithm has been used as a training algorithm using the concept of information entropy to classify repeating patterns. The database consists of 12 characteristics such as age, gender, history of blood pressure, etc. The results showed that the designed prediction system has the ability to predict heart attack with high accuracy, but the high speed has been ignored in this research.

In [4], in order to improve the heart disease prediction system, cases based on measurements taken by echocardiography tests and knowledge discovery methodology were used in the database. He reviewed data from 7,008 patients between 2008 and 2010 for 20 variables that were narrowed down to 15 by an expert. Then he investigated the available algorithms in DM using WEKA software and finally he chose decision tree algorithms, NN and Bayesian clustering algorithm. All tests were divided into two groups. The first group with 15 features and the second group with 8 features. The obtained results show that despite the good performance of all models, the J48 algorithm implemented on 8 features was more efficient with 95.56% accuracy, that this research only examined the accuracy of the algorithm and the speed of the algorithm was not considered important Also, each algorithm was evaluated alone, and the combined method could provide better results.

In [5], the researcher investigated and compared DM algorithms and by choosing disease diagnosis as the target variable and 7 predictor variables of different types of age, type of chest pain, resting blood pressure, blood sugar, and resting ECG, the maximum heart rate and angina pectoris caused by exercise were observed, and variables such as angina pectoris caused by exercise, ECG while resting, age, type of chest pain, blood pressure at rest, blood sugar are known as important variables that need attention. These factors will reduce this disease. The obtained results showed that decision tree and CHAID algorithm and simple Bayes network provided acceptable results with high accuracy. In this research, the focus was on finding the most important risk factors for heart disease and also on the accuracy of the algorithms, but the response speed of the algorithms has been ignored.

In [6], only eleven characteristics of South African patients' dataset, such as age, sex, chest pain, etc., have been investigated. WEKA DM software has been used for its effectiveness in discovering, analyzing and predicting patterns. The researcher has used J48, Bayes Net (BN), NB, Simple Cart, REPTREE algorithms to classify and improve a model for heart attack diagnosis. The research results did not show a significant difference in the use of different algorithms. The prediction accuracy found in the J48, Simple Cart and REPTREE algorithms suggests that the parameters used to predict heart disease are reliable. The results obtained in this research did not advance the problem of predicting heart disease and were only a repetition of the results of past research. Also, more focus was on the parameters used in the prediction of heart disease.

In [7] researchers presented an effective association classification algorithm using a genetic approach. The main motivation of using genetic algorithm (GA) in discovering high-level prediction rules is that the discovered rules are understandable and have high prediction accuracy. Six datasets from SGI database and 2 datasets from UCI have been used for this research. At first, the Gini coefficient was obtained for different datasets and the features were selected based on this index. Then the accuracy was obtained using the 10-fold cross validation method. Then it was compared with Genetic Network Programming j48, Naive Bayes, Neural Network algorithms. The results show that most classification rules help in better prediction of heart disease. In this test, the proposed algorithm provided an average accuracy of 88.9. Since the compared algorithms often have relatively equal results with the presented method, it seems that the presented method does not have a significant advantage over the previous methods.

In [8], the pre-processed data are clustered using a combination of two common KM and Particle Swarm Optimization (PSO) algorithms. Also, the MAFIA algorithm is used to explore the maximum repeating patterns. Repeated patterns can be classified using C4.5 algorithm as a training algorithm. Although the PSO algorithm is a good algorithm, it is not suitable for large and complex datasets. In the initial stage, the PSO clustering algorithm is executed to find the location of the cluster center. After 10 runs, these locations have been used as the initial center of the K-means algorithm in order to refine and produce the best clustering solution. MAFIA algorithm is then used to integrate a depth traversal (DFT) with effective pruning mechanism. In order to create classification rules, the decision tree and C4.5 algorithm have been used. By combining these two algorithms, the accuracy of the K average algorithm has improved and this combination has provided an acceptable accuracy.

In [9], a classifier method for heart disease diagnosis using NB algorithm is presented. Medical data are divided into 5 categories named (negative, mild, average, high, very high). Also, if an unknown sample is entered, the system predicts the label of the sample category. Therefore, two main functions called classification (training) and prediction (testing) are implemented. 14 parameters such as age, gender,

cholesterol, chest pain, etc. are checked. The 303 records were evaluated in the training phase and 276, 240, and 290 records were evaluated in the testing phase of the dataset, and the results of the analysis performed on the Cleveland dataset showed that the accuracy of the proposed method was 88.76, 89.58, and 88.96%, respectively.

Among the works that have been done in this field, it is mentioned in [10]. Researchers presented a high-accuracy hybrid method for diagnosing heart disease. The presented method is able to increase the performance of the NN by about 10% by increasing its initial weight by using the GA, which is a better weight for the neural network. Alizadeh Sani's z-dataset includes the information of 303 patients, 216 of whom suffer from CAD. ‹The 54 characteristics were collected for each patient, and these characteristics were divided into the following 5 groups: 1- Demographic Characteristics such as age, sex, etc., 2- Symptoms and results of clinical tests such as blood pressure, type of chest pain and so on. 3- Electrocardiography, 4- Echocardiography and 5- Laboratory tests. To select the characteristics of four famous ranking methods of Gini coefficient, weighting by SVM, information acquisition and principal component analysis were considered. With this methodology, their results showed that the rate of accuracy, sensitivity and specificity on Alizadeh Sani's z-dataset is 93.85%, 97% and 92%, respectively.

The main idea considered in this research is the integrated use of neural networks based on Meaning fuzzy clustering to improve performance. First, the training data is clustered using the CMF algorithm (the centers of the clusters are calculated). Fuzzy clustering is sensitive to noisy and outlier data and to increase the possibility of selecting data that are misclassified. The cluster centers should be scattered as much as possible, and using this type of clustering reduces the amount of training data and increases the speed of training. The CMF algorithm is repeated several times and each time it creates different input data with different degrees of membership (which is used to have better input characteristics to increase accuracy). Then the neural network algorithm is implemented and the combination of these two algorithms will increase the accuracy in heart disease diagnosis. In order to analyze the proposed method, MATLAB software is used, and UCI standard dataset is used to evaluate the parameters.

3. THE HEART DISEASE

3.1. Risk factor for heart disease

Evidence points to a number of risk factors for heart disease: age, sex, high blood pressure, blood lipids, diabetes, tobacco smoking, processed meat consumption, excessive alcohol consumption, sugar, family genetics, obesity, lack of physical activity, psychosocial factors and air pollution. While the contribution of each risk factor is different between different communities or ethnic groups, the overall contribution of these risk factors in epidemiological studies is significant. Some of these risk factors, such as age, gender, or family history, are immutable. But many of these important cardiovascular risk factors can be modified by lifestyle changes, social changes, drug treatment, and prevention of high blood pressure, lipids, and diabetes.

3.2. Hear disease diagnosis methods

Most of the time, the treating doctor does a lot of investigations to get enough information to diagnose your disease and make the right decision about the appropriate treatment. The set of tests that your doctor requests for you may not be necessary in another person with the same condition [11]. The results of a survey may provide information that reveals the need for more surveys to complete the information. You may also need to repeat the same test several times to determine exactly how your heart is responding to different medications, surgical procedures, or other treatments. Medical history is also a determining factor in requesting appropriate examinations [12]. There are several methods to diagnose heart disease, which we will discuss below.

4. THE PROPOSED METHOD

4.1. The used dataset

Research methods refer to the methods of designing research studies and data analysis procedures. Achieving the goals of the research is not possible unless the research process is carried out with the correct methodology. In general, the research method can be defined as a set of reliable and systematic rules, tools, and ways that are used to investigate facts, discover unknowns, and achieve solutions to problems. In this section, the research method, the statistical population, the collection method and the conceptual and operational definition of the variables have been discussed. Scientific researchers can be divided into three basic, applied and developmental researches based on the purpose [13]. Also, in the classification based on the method, researches are divided into historical, descriptive, correlational, experimental and causal.

The intended method for this research is cross-sectional. This research seeks to discover and examine the relationships between factors and specific conditions or the type of event that existed or occurred, through the study of data and their results. In other words, it seeks to investigate the possibility of the existence of cause and effect relationships by observing the existing results and their previous context in the hope of finding a diagnosis and predicting the occurrence of the phenomenon.

The proposed methodology consists of different parts. The heart database section includes traits that are used to distinguish sick from healthy individuals. As

stated earlier, the database consists of 14 columns and 125 rows. The 8 columns represent the attributes and one column represents the class label. The statistical population consists of all the elements and people who have one or more common characteristics on a geographical scale (global or regional). The statistical population of this research is cardiac patients in two hospitals from the UCI heart disease database [14]. The data used in this research was extracted from the machine learning theory database of the University of California at archive.ics.uci.edu/ml. This dataset is related to chronic heart diseases and the variables in it are explained in Table 1.

Table 1. Variables in the dataset.

Description	Variable type	Variable name	Row
Patient age in years	Numerical	Age	1
Indicates the upper limit of systolic blood pressure in mmHg	Numerical	Blood pressure	2
One of the numbers (1/005,1/010,1/015,1/020,1/025) obtained from the measurement of different components of urine includes waste materials and solutes. High specific gravity indicates concentrated urine. This number is used for concentration and secretion power. The normal range is between 1.005 and 1.02.	nominal	Specific density	3
It is one of the numbers [15] [1] [2] [16] [3]. Albumin is a type of protein in the blood that is filtered by the heart. Excreting more than 3.5 grams of albumin in 24 hours through urine is a sign of a kind of kidney disorder called Albuminuria.	nominal	Albumin	4
It is one of the numbers [15] [1] [2] [16] [3]. Glucose is the sugar present in the blood and is considered the main food of cells. Normally, there are very small amounts of glucose in urine, but under normal conditions, it is less than 4 grams.	nominal	Sugar	5
Variable zero and one indicating normal or abnormal number of red blood cells in urine. Excessive presence of red blood cells in urine is known as hematuria.	binary	Number Red globules	6
indicating the presence or absence of infectious cells in the urine	binary	Pus cells	7
Observing or not observing the infectious mass in the urine	binary	Pus cells clump	8
Observing or not observing bacteria in urine	binary	Bacteria	9
It shows blood sugar regardless of when the food was consumed. Several random checks may be made throughout the day. Because the amount of blood sugar in healthy people does not change much during the day. If the blood sugar changes a lot during the day, it probably indicates a problem with the heart.	Numerical	Random blood glucoses	10
Indicates the amount of urea in urine	Numerical	Urine	11
Creatinine is a product of muscle metabolism that is excreted from the heart. When the level of creatinine in the blood is higher than the normal level, it is a sign of heart failure, and therefore it is a good measure to evaluate the heart's function.	Numerical	Creatinine	12
Excess excretion of sodium by the heart indicates a disorder called hyponatremia.	Numerical	Sodium	13
Excess excretion of potassium by the heart indicates a disorder called renal tubular acidosis.	Numerical	potassium	14
If there is hemoglobin in the urine, it can be a sign of kidney dysfunction	Numerical	Hemoglobin	15
According to the percentage of red cells in the blood, it is called hematocrit or sea blood. For example, if a person's hematocrit is 40, it means that 40% of the blood volume is made up of red blood cells and the rest is plasma. The average hematocrit of men is 42 and the average hematocrit of women is 38.	Numerical	Hematocrit	16
The normal amount of white blood cells is 4,500 to 10,000 white blood cells per microliter.	Numerical	Number White globules	17
The number of red blood cells in a normal state is 3.6 to 5.5 million in the blood of women and 1.4 to 6 million per microliter in the blood of men.	Numerical	Number of red blood cells	18
It indicates the presence or absence of blood pressure in the patient.	binary	Hypertension	19

Description	Variable type	Variable name	Row
Indicates the presence or absence of diabetes	binary	Diabetes	20
It shows the presence or absence of heart disease. In coronary heart patients, the coronary arteries are narrowed (Stenosis) and the heart muscles are deprived of sufficient blood and oxygen.	binary	Coronary artery disease	21
indicating good or bad appetite	binary	Estenosis	22
indicating the presence of swelling in the area of the back of the leg. Excess excretion of protein by the kidney leads to this complication.	binary	Estenosis	23
It indicates the presence or absence of anemia in the patient.	binary	Estenosis	24
Indicates the presence or absence of heart disease	binary	output (handle)	25

4.2. The pre-processing stage

Data pre-processing will be done in two parts. The first part includes completing the missing values and the second part also includes detecting outliers and removing them, which can affect the final results.

In general, the methods of filling missing values are:

Delete data. This means that data with missing values are completely removed from the data set. It is obvious that this method is used when a small amount of data has missing values. Because if we remove too many from the set, the remaining data will probably cause overfitting. This means that the accuracy of the model will be high on the training data and low on the test data.

Considering a global value for all missing values. In this case, a fixed value is considered for each attribute (variable) and all missing values are replaced with that number. This task is usually used when a constant value can be determined for all missing items [17]. For example, if the data is collected from a place such as a university and most people's jobs are registered as students, the same amount of students can be determined for the missing job values.

Using the mean value of the variable for all missing values. where for each variable, the average value is calculated using the available data and is replaced by all the missing values of that variable. This method may be sensitive to outlying and outlier data [18]. In addition, in cases where there is a nominal variable (such as hair color or type of car, etc.), in this case it is not possible to use this method.

Using the median value of the data for missing values. Since the average is sensitive to outliers and there may be outliers in some data sets, in some cases where there is a possibility that the average is inappropriate, the average of the data is used to determine Missing values are used. Of course, if the distribution of the data is symmetrical and there are no outliers, using the median or mean will have almost the same result [19]. Because in such a case, the mean and the average are close to each other and therefore the estimated number for the missing in these two methods are almost equal.

Determination of missing values using learning methods. In these methods, the variable whose value is missing is selected as the dependent variable and other variables will also play the role of the independent variable [20]. Then, using a learning method such as linear regression (which does not threaten the risk of overfitting), the missing value is predicted.

The fifth method is more accurate than the other four methods. Because outlier data, asymmetry of data distribution, etc., which strongly affects methods 1 to 3, shows a lesser effect in the fifth method. On the other hand, using the fifth method will also have the risk of overfitting. Because the use of a complex method (for example, a neural network with a large number of hidden layers) may perfectly match the training data, but have very low accuracy and precision in the model testing phase [21]. Usually, complex nonlinear models are not used to overcome the risk of overfitting. For this reason, multiple linear regression will be used to estimate the missing values.

Linear regression is a relationship between a dependent variable and a number of independent variables based on which the extracted equation can be used to fill in the missing values. The general form of a regression equation is as $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ In which Y Dependent variable vector and $X = [x_1, x_2, ..., x_n]$ Independent variables are the objective of regression to find coefficients $\beta_0, \beta_1, ..., \beta_n$ is.

For this purpose, the column related to the variable whose missing values are to be predicted should be considered as dependent variable and other variables as independent. With such an action, it will be possible to estimate the coefficients using relation 1.

$$\beta = \left(X^T X\right)^{-1} X^T Y \qquad (1)$$

Detection of outliers: Outliers usually refer to data that are much smaller or larger than other members in a set. The existence of these data can affect the results and actually damage the accuracy of the results. Therefore, it is necessary to analyze such cases and remove them from the data if they exist. These data can be the result of one of the following:

1)Inaccuracy of the observed measurement

2)Collecting data from different communities

3)Measurement for a rare incident or event

4)Excessive skewness of data in frequency distribution

Outlier data detection methods are classified into two

groups, univariate and multivariate, based on the number of variables. This means that only one variable is considered in the analysis of outlier data and the outlier value is determined for it (univariate methods) or that the variables in the data set are considered and checked simultaneously (multivariate methods)) [22]. One-variable detection methods are among the simplest methods available in the literature. Among the univariate methods, we can mention the median technique, box plot, smoothing methods such as containerization, etc. Among the multivariate methods, we can mention multiple regression, NN, distance, and etc.

Some of the existing methods in the field of outlier data detection are as follows:

Grubb's test: This test was presented by Grubb [23] to determine outlier data and is based on statistical methods. In Grubb's test, it is assumed that the data follows a normal distribution. This method is iterative and outliers are identified and discarded in each iteration. This algorithm continues until no extraneous data is detected. The statistics of Grub test are as follows:

$$G = \frac{\max_{i=1,\dots,N} |X_i - \bar{X}|}{s} \qquad (2)$$

where in $\bar{X}$ and s The mean and standard deviation of the data are respectively N is also the number of data. In each iteration of the algorithm, the value G calculated and if its value is greater than the relation 3, the largest and smallest data are removed from the data.

$$\frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\frac{\alpha}{2N}, N-2}}{N-2+t^2_{\frac{\alpha}{2N}, N-2}}} \qquad (3)$$

where in $t^2_{a,b}$ A point of distribution t With b It is a degree of freedom that has equal probability on the right side a is also α The significance level of the test is usually 0.05.

Bonferroni method: In this method, the median of the data is first calculated and then the standard deviation is added and subtracted from it. With this action, a range will be formed, the data outside of which are outliers and will be removed from the calculations.

Clustering method: In this category of methods, the number of optimal clusters is first determined by using the existing indicators, and then the data is assigned to the clusters. It is believed that the cluster with the lowest density contains outlier data. Therefore, these data are removed and the analysis is performed on the remaining data.

NN: NNs are one of the computing methods that, with the help of the learning process and using processors called neurons, try to provide a mapping between the input space by recognizing the inherent relationships between the data. The hidden layer or layers processes the information received from the input layer and provides it to the output layer. Each network is trained by receiving examples. Education is a process that ultimately leads to learning. Network learning is done when the communication weights between the layers change so that the difference between the predicted and calculated values is acceptable. After learning, the network can decide whether a particular data is an outlier or not.

Besides the introduced methods, there is another technique called CM and single class NNs. In general, this method is used to solve a problem in which one category exists as the target category and the rest of the categories are known as outlier data. The main purpose of this method is to find a boundary around the data related to the target category, which can be used to identify outliers. Fig. 1 shows a boundary found by CM and single-class NNs.

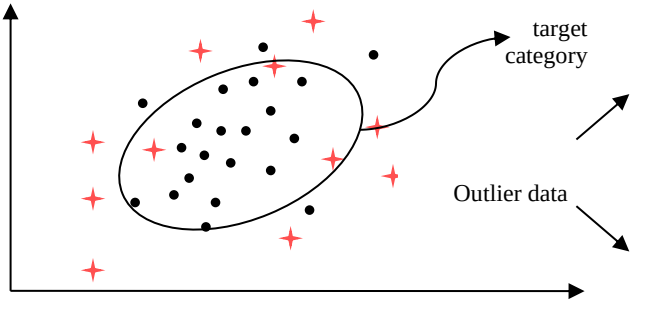

Fig. 1. CM and single class NNs.

Since this method is a non-parametric technique and does not require the assumption of normality of data distribution, single-class version of CM and NNs will be used to discover outliers. The mentioned model is in the form of relation 4:

$$\begin{aligned} \min \quad & R^2 + \frac{1}{\upsilon m} \sum_{i=1}^{n} \xi_i \\ & \left\| \phi(x_i) - c \right\|^2 \le R^2 + \xi_i \\ & \xi_i \ge 0 \end{aligned} \qquad (4)$$

where R ξ_i and c are the decision variables of the model and represent the radius, deviation of the data from the coverage area and the center of the coverage area, respectively. In addition, the size of the trade-off between the coverage radius and the number of data outside the coverage area is determined by $\upsilon \in \{0,1\}$ the parameter. So that if the value of this parameter is close to zero, more importance is given to the points outside the coverage and therefore the value becomes R larger. Also, if υ it goes to one, the importance of the data outside the coverage decreases and the value of the coverage radius decreases. The best parameter υ value

is the value that balances the trade-off between these two variables [24] [21].

The validation method: In general, validation methods are divided into two parts: Exhaustive validation techniques and Non-Exhaustive validation methods. In full validation methods, the number of k data from the available N data is used for testing and the rest for training, and this is done for all modes. In other words $\binom{N}{k} = \frac{N!}{k!(n-k)!}$ it is executed as many times as the model. It is quite clear that the number of times the model is run for these methods is very high and therefore limits their use. On this basis, incomplete validation methods are usually used, of which the 10-layer validation method is one of them.

In addition to these methods, there is another method called bootstrap, which randomly selects data for machine training and can also select a duplicate sample. In fact, sampling is done by placement to select the training set.

In line with the implementation of the proposed plan, the data at hand have been divided into two parts, training data and test data, using the 10-layer cross-validation method, and after 10 times of training and testing the model with Different layers, the best amount of indicators have been reported.

4.3. The method stage

The CM clustering is an unsupervised learning method that separates the data into different parts, which is based on a similarity criterion and is done in such a way that the members in the cluster are the most The similarity and the members between the clusters have the least similarity. For a more technical explanation, suppose the clusters are represented by symbols C_i and the center of each is also represented by c_i symbols. so $i = 1,2,\dots,c$. that Non-fuzzy clustering methods assign each data exactly to one cluster, which may cause inappropriate clustering. Because a data may be located in a point between two or more clusters and by applying this restriction that each data belongs to a cluster, the similarity of the mentioned data with other clusters will be ignored. But on the other side, there is fuzzy clustering in which each data can belong to different clusters with different degrees of membership. If it indicates the degree of membership of the data x_j to the cluster C_i, then we will have fuzzy $u_{ij} \in [0-1]$ clustering in which the value closer to 1 will indicate the greater belonging of the data to that cluster. Therefore, the degree of membership of a data for all clusters is defined as relation 5.

$$u_j = (u_{1j}, u_{2j}, \dots, u_{cj}) \qquad (5)$$

If a data set n contains data, the matrix of membership degrees will appear as a matrix $c \times n$ and according to the relation 6.

$$U = (u_1, u_2, \dots, u_n) \qquad (6)$$

The following restrictions are in place for the matrix of degrees of membership:

$$\sum_{j=1}^{n} u_{ij} > 0 \forall i \in \{1,2,\dots,c\} \qquad (7)$$

$$\sum_{i=1}^{c} u_{ij} = 1 \forall j \in \{1,2,\dots,c\} \qquad (8)$$

The first constraint requires that all clusters have at least one member, which means that no cluster is empty. In addition, the second limitation also states that the total membership degrees of a data on all clusters is equal to 1. These two restrictions make a data not belong to only one cluster. Each type of clustering method (fuzzy or non-fuzzy) uses a fitness function, which is actually a criterion for clustering. This function is a loss function and therefore a proper clustering minimizes its value. With F the notations made so far, this function is shown by and defined as follows

$$F = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \left\| x_j - c_i \right\|^2 \qquad (9)$$

where the value is m greater than 1 and is called Fuzzifier exponent. Fuzzy clustering uses an iterative method to solve the fuzzy clustering problem. In the mentioned method, the degree of membership and the center of the clusters are obtained from the following relations in an iterative procedure:

$$u_{ij} = \frac{1}{\sum_{l=1}^{c} \left(\frac{d_{ij}^{2}}{d_{lj}^{2}} \right)^{\frac{1}{m-1}}} \qquad (10)$$

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \qquad (11)$$

To perform fuzzy clustering, the following steps are taken:

A) The initial value of the fuzzification power (greater than 1) and the U matrix are chosen randomly.
B) The center of the clusters is calculated using relation 11.
C) After calculating the centers of the clusters, the U matrix is updated.
D) Steps (b) and (c) are repeated until the stopping condition is met. The stopping condition is usually no change in the centers of the clusters or the value of the fitness function

The proposed hybrid model includes the use of CM clustering in the neural network structure. In fact, the new model aims to improve the final prediction results by using an unsupervised learning algorithm in the structure of a supervised algorithm. This is done in such a way that the neural network also uses the information related to fuzzy clustering. Based on this, two approaches are presented to perform the composition.

In the first approach, which is called membership

degree method, at first the data is clustered in fuzzy form using CM algorithm and then the membership degrees obtained from the clustering process are given as input to the NN and after pre-processing. The nose of the output values will be compared against the previous methods.

In the second approach, which is named as clustered data, after the fuzzy clustering process, the main data along with membership degrees will be used as input for the NN.

In order to explain further, suppose that X represents the data of the problem and Y is its corresponding output. By performing fuzzy clustering on the problem data, MD is also the degree of membership related to the clusters. In this case, Fig. 2 will show the two proposed approaches for the hybrid neural network-fuzzy clustering model.

In better words, fuzzy clustering will lead to the creation of two modified data sets from the original data set, one of which includes degrees of membership and numbers of categories, and the other includes degrees of membership, variables It includes the original and the number of the categories. To perform fuzzy clustering, iterative method according to relations 3-10 and 3-11 is used and it will converge to the optimal solution by using an innovative algorithm whose steps are in accordance with the following clauses.

1. An initial population of solutions with size P is generated. Each answer is a vector containing membership degrees related to the data.
2. For each answer, the value of prediction accuracy is calculated and produced by name$f_i, i = 1,2,\dots,P$.
3. Answers are sorted according to accuracy and in descending order.
4. An answer is selected to produce a new answer. Each of the answers has a certain chance to be selected, and this chance will be calculated using the relation 12.

$$V_i = \frac{2(p-i+1)}{p(p+1)} \qquad (12)$$

5. After selecting the ith answer, a new answer will be generated with a normal distribution with a mean vector equal to the membership degrees of the ith answer and a variance of 1.
6. The new answer is replaced in the existing population.

The introduced steps continue until a certain number of repetitions until the best answer is obtained and reported.

The code network of the proposed method is shown in Fig. 3.

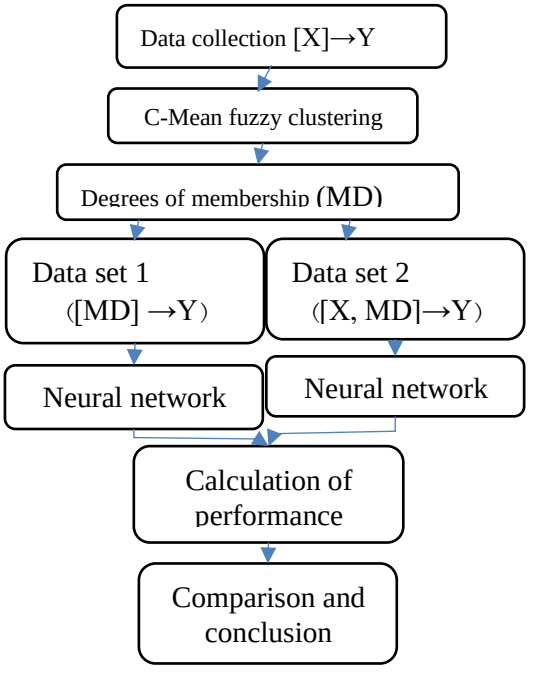

Fig. 2. How to combine CM techniques and NNs.

Initialization
Impute missing values
Detect and remove outliers (One-class C-Means)
Clustering Phase
For different number of clusters
Do
Cluster data using fuzzy C-Means
Compute objective function of clusters which shows clustering efficiency
Select the best number of clusters (lowest objective function)
End For
Classification Phase
Construct new datasets
Classify data using Neural Network
Analyze
Compute measures
Report measures
End

Fig. 3. pseudocode of the heuristic algorithm.

Performance indicator: To measure the performance of the proposed model and compare it with the previous methods, the accuracy index will be used, whose value is equal to the total number of correctly labeled data on the total available data. Using this index can provide the possibility of comparison.

Paired t-test will be used to confirm the superiority

of the proposed method over the traditional method. This test includes the following steps:

The value of the index difference for different models is calculated and named as D_i. The mean and standard deviation of the difference values are called S_D S and $\bar{D}$. The test statistic is determined as $t = \frac{\bar{D}}{S_D/\sqrt{n}}$, where n is the number of tests.

The critical value of the test is obtained from the t distribution tables and is equal to $t_{\alpha,n-1}$. where α is the significance level and $(1-\alpha)$is the confidence level of the test.

The decision is made in such a way that if the value of the statistic is greater than the critical value of the test, it means that the proposed method has performed better than the traditional method.

5. RESULT AND DISCUSSION

This section of the research will deal with the implementation of the proposed approach introduced in the previous section. Based on this, at first, the missing values in the data are estimated by multiple linear regression method and then they are analyzed with CM and single-class NNs in terms of outlier data. After identifying and removing these values, if any, the data is divided into different clusters using the fuzzy clustering method that was introduced in section 2-3 and classified into two types with the proposed combined method. They are classified. The obtained results will be used to compare the proposed method and the traditional method.

5.1. General calculation results

All the calculations in this section have been done in the MATLAB R2015b software environment. A 10-layer cross-validation method was used to determine the training/test data set, and based on this, each model was run 10 times and the average value of the performance indicators was reported. The stopping condition of the algorithm is set in such a way that if no improvement is observed in 20 consecutive iterations, the clustering is stopped and the best clusters obtained that correspond to the lowest value of the objective function in relation 5 are selected. They will be stored and used in the next steps.

5.2. Completing the missing values and detection of outlier data

In accordance with the content presented in the third section, the completion of the missing values has been done using the multiple linear regression method. This method has indicators called Coefficient of determination and Adjusted coefficient of determination, which are indicated by symbols R^2 and R^2_{adj} respectively. These indices are between zero and one and their closeness to the number 1 shows the higher efficiency of the model. The mentioned indicators are defined as follows:

$$R^2 = 1 - \frac{SSE}{SST} \quad (13)$$

$$R^2_{adj} = 1 - \frac{SSE/_{N-n-1}}{SST/_{N-1}} \quad (14)$$

where N represents the total number of data and n represents the number of independent variables. On the other hand, SSE and SST are also Total sum of squares (SST) and Error sum of squares (SSE) and are calculated with relations 15 and 16, where the actual value of the dependent variable, its average value and It is also the predicted value for it.

$$SSE = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (15)$$

$$SST = \sum_{i=1}^{N}(y_i - \bar{y}_i)^2 \quad (16)$$

In fact, the sum of squared error indicates the sum of the error values in the prediction of the output variable, and the total sum of squares also indicates the dispersion of the actual output values. The lower the ratio of SSE to SST means the better performance of the model, and therefore its difference from the number 1 known as R^2 is an indicator to measure this fact. In other words, this index indicates the probability of correlation between input and output variables. This coefficient actually expresses the approximate results of the desired parameter based on the defined mathematical model that matches the available data. In addition, the R^2_{adj} index also divides the normalized values of the sum of squares in order to eliminate the effect of the number of variables used.

Based on the analysis performed on the data set which includes 400 records and 25 variables, the number of variables with missing values is as described in Table 2.

In addition, the frequency of missing values is also summarized for each of the variables according to Table 3.

Table 2. Frequency of missing values for each of the data.

Abundance	Number of missing values for each data
158	0
45	1
33	2
37	3
31	4
33	5
12	6
20	7
8	8
12	9
4	10
7	11
400	Total

Table 3. The frequency of missing values for each of the variables

Frequency of missing data	Variable number
9	1
12	2
47	3
46	4
49	5
152	6
65	7
4	8
4	9
44	10
19	11
17	12
87	13
88	14
52	15
71	16
106	17
131	18
2	19
2	20
2	21
1	22
1	23
1	24
0	25
1012	Total

The total of 1012 obtained in Table 3 shows the total number of missing persons. Using the multiple linear regression method, these values were completed and the indicators of the coefficient of determination and the modified coefficient of determination are summarized in Table 4.

Table 4. Coefficient of determining regression in completing missing values (%)

Modified coefficient of determination	The coefficient of determination	Variable number
95/07	94/83	1
91/58	91/35	2
46/52	46/39	3
60/84	60/67	4
88/41	88/16	5
69/46	69/18	6
78/90	78/67	7
93/30	93/07	8
96/80	96/56	9
82/62	82/38	10
85/73	85/51	11
91/24	91/00	12
86/16	85/88	13
99/60	99/28	14
40/99	40/87	15
45/16	45/02	16
94/94	94/62	17
51/31	51/12	18
65/49	65/32	19
69/19	69/01	20
94/70	94/47	21
86/69	86/47	22
88/04	87/82	23
89/67	89/45	24
-	-	25

Outlier data are determined using CM method and neural networks in such a way that maximum 20% of the data are removed from the calculations. For this purpose, different values have been considered for the adjustment parameter υ and based on that, the number of outlier data in each state has been determined. Table 5 shows the results of this processing. In this regard, it is clear that for different values of the $\upsilon \in \{0,1\}$parameter, none of the data are identified as outliers, and therefore, the existing data whose missing value is predicted in the previous step are used for implementation. The proposed method will be used.

Table 5. The number of detected outlier data for different values of the setting parameter

Number of data outliers	parameter value υ
0	0/1
0	0/01
0	0/001
0	0/0001
0	0/00001

5.3. Solving the proposed model

In order to implement the proposed hybrid model according to Fig. 2, the completed data in which the outliers have been examined will be used. Based on this, fuzzy clustering is done using the heuristic algorithm of section 6-3 and then both approaches related to Fig. 3 will be applied to it.

5.4. Data clustering

In order to present the first report in the computational results section, Fig. 4 shows the values related to the objective function in successive iterations of the algorithm. The relevant clustering was obtained using a population size of 50 and for 4 clusters. Also, the power of fuzzification is set equal to 5 in this algorithm. It is clear that the value of the objective function has almost converged after 6000 iterations, and its changes in successive iterations are not significant. However, the stopping condition occurred approximately in the 17000th iteration and therefore the proposed algorithm was stopped at that point.

From the point of view of comparing the values of the objective function and determining the best number of clusters using the fuzzification power of 5, the clustering process was performed 10 times for the number of clusters from 2 to 6 and then the confidence interval using software MINITAB software has been reported for the value of the objective function of each of the states. Table 6 is the summary of this experiment.

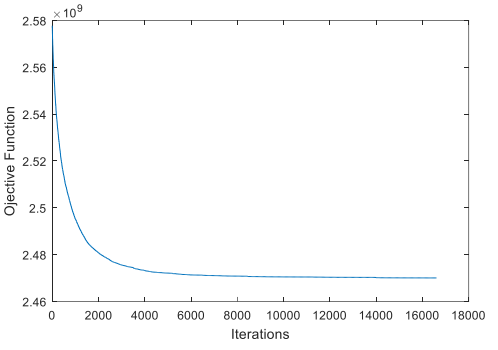

Fig. 4. The value of the objective function of fuzzy clustering in successive iterations of the heuristic algorithm

Table 6. The confidence intervals of the values of the objective functions of clustering according to the number of different clusters

6	5	4	3	2	Number of clusters
2534061804	2544815328	2470446082	2444070668	2580102378	Average
2534061785	2544815133	2470445901	2444070531	2580102158	At least
2534061842	2544815572	2470446224	2444070845	2580102524	Maximum
2534061791	2544815212	2470445995	2444070594	2580102305	95 % interval estimate
2534061817	2544815443	2470446170	2444070742	2580102470	

D.1. Making a bundle

In this subsection, according to Fig. 2, the proposed classification models will be created. In this regard, the indicators of accuracy, sensitivity and transparency will be used to check the performance of the models. For this purpose, at first, only the membership degrees were used as the input of the NN, and after 10 times of running the model, the average values for these indicators were calculated. Table 7 summarizes the relevant results. Considering that in the previous subsection, the number of 3 clusters had the best performance value from the point of view of the objective function value, so these 3 clusters will be used to build the data set related to the proposed categories.

Table 7. performance indicators obtained from the combined model

Transparency	Allergy	Precision	Proposed model
79/98	06/96	01/97	Conventional model
91/96	29/95	79/96	
39/99	75/96	79/98	
89/95	21/93	21/95	
91/55	11/53	01/55	
00/79	95/57	66/65	Combined model 1
00/79	94/57	66/65	
00/79	57/61	33/68	
90/70	73/74	33/71	
27/82	31/56	83/65	
100	36/97	33/98	Combined model 2
100	05/96	50/97	
79/99	67/98	13/99	
63/88	100	83/95	
100	94/28	00/55	

In this case, it is clear that the first proposed model, which includes the use of clustering results as the input of the category model (proposed model 1), is not very accurate. However, the second model, which includes the combination of the original data and the clustering results, has a higher accuracy, and in the meantime, by using the number of 3 hidden layers, each of which contains 2 neurons, the accuracy is 13/%. 99 has also been achieved.

The comparison between the proposed models and the conventional model shows that the conventional method is superior to the first proposed model; However, the second proposed model is superior to both mentioned models and has achieved better performance indicators.

5.5. Comparison of the proposed method with the conventional method

In order to compare the second proposed method and the traditional method, the paired t statistical hypothesis test has been used to confirm the superiority of the proposed method with more certainty. In order to implement this test, the results of table 7 for the conventional machine and the second proposed machine have been used. After performing this test, the results are shown in Table 8.

Table 8. Test results related to the comparison of the proposed method with the traditional method

Result	$P-Value$	Degree of freedom	The standard deviation	Test statistics
The superiority of the proposed method	0/000	4	0/4924	-2/7056

After the test, a decision is made about it using the $P-Value$ index. Thus, if the value of $P-Value$ is less

than 0.05, the superiority of the proposed method can be confirmed. Based on the obtained results and considering that the $P-Value$ value in this experiment was equal to 0.000, the superiority of the second proposed method over the conventional method can be confirmed.

5.6. Analysis of the effect of population size

All the data obtained in Sect. 5 were obtained using a population size of 50, and it is possible that using a higher population size may lead to better results. In order to check the effect of the population parameter of the algorithm on the accuracy of the model, other values were also used and the results are shown in Fig. 3. (Table 9)

Table 9. Accuracy of the model for different values of the population size

Accuracy of the model	The value of the objective function	Population size	Row
99/07	2451925691	30	1
99/13	2444070668	50	2
99/18	2414810368	70	3
99/18	2385710423	100	4
99/31	2351053187	150	5
99/33	2310473621	200	6
99/35	2285791322	500	7

Based on this, as the population size increases, the value of the objective function related to clustering will decrease. Such a reduction means an improvement in data clustering, and subsequently, with the improvement of clustering, the accuracy of the classification model is also improved. In this regard, the best accuracy will be related to the population size of 500 with 35.99%.

This section of the research was dedicated to the implementation of the proposed model on the events related to heart disease. In this regard, the data used in the UCI Repository database were collected and pre-processed. For this purpose, the missing values among the completed data and the outliers in them were also evaluated using CM method and single class neural networks. Then, using the CMF clustering algorithm, the existing data were clustered and based on two approaches, they created a new data set for classification. In the first approach, membership degrees related to the data were assigned to the category model as input. In the second approach, the existing dataset was considered as the input of the model along with membership degrees. According to the calculations, the second approach has the advantage of accuracy compared to other models, and this fact has been confirmed by using the statistical assumption test. Finally, the analysis performed on the population size shows that the accuracy of the proposed model is improved by increasing the mentioned parameter and will increase up to 35.99%.

6. CONCLUSION

The current research has presented two new approaches for predicting heart disease. These two approaches use fuzzy clustering to modify the heart disease dataset. In the first approach, membership degrees are used instead of the existing data set to perform the classification process. In this way, the data clustered and then the membership degrees are given as input to the NN and the main attributes are removed. The second approach also suggests the use of membership degrees in addition to the attributes in the data set. As a general summary of the different sections of the research is of concern:

- At first the current authors discussed the generalities of the upcoming problem and explained the necessity of doing it.
- Then the current authors also devoted to thematic literature review. For this purpose, articles related to the field of medical diagnosis and the use of DM tools for diagnosis have been introduced and the recommended methods have been described in them.
- After it, the current authors are dedicated to the introduction of combined approaches to perform prediction in the field of heart disease. In this regard, two approaches have been presented to combine the fuzzy clustering method and the DM.
- Finally, the introduced approaches have been evaluated numerically, and the accuracy, sensitivity and transparency indicators have been reported for each one. The statistical hypothesis test has determined that the second approach, that is, the combination of degrees of membership and attributes in the data set of cardiac patients, can improve the accuracy of the model compared to the conventional mode.

6.1. Research results

The heart is a vital organ for the body, and the diagnosis of heart disease and the subsequent provision of appropriate treatment methods are considered an important measure in the field of medicine. Considering the abundance of information in the field of medicine, DM can increase the quality of services provided to patients and reduce treatment costs by extracting the knowledge hidden in the heart of this information and using it in performing medical processes.

DM can help doctors in predicting and diagnosing the appropriate treatment method for heart patients by

extracting knowledge from a large amount of medical information.

Early diagnosis of any type of disease is considered an essential factor in treatment. In this research, our goal has been to design a system that can help doctors in diagnosing diseases. This research introduces a diagnostic CM assisted neural networks and decides which technique is useful in heart disease diagnosis.

In this research, the current authors used the database obtained from the machine learning laboratory at UCI University. All patient data were trained using neural networks. Choosing the best value of parameters for a certain kernel is a fundamental problem in the NN method, which can be used to diagnose a common disease with simple clinical criteria. In this research, the combination of neural networks and CM algorithm was used. The results show that the above combined algorithm works better than other previously performed algorithms and the detection percentage is much higher than the work of others in this field. Performance parameters, such as classification accuracy, sensitivity and specificity of NNs and RBF have been high, and therefore we have presented a method that can be a suitable option for the classification process.

The results obtained from the numerical tests have determined that the use of membership degrees obtained along with the attributes in the data set can improve the accuracy of the model. In this regard, the second approach by performing such a combination has been able to increase the accuracy of the model to 99.13% using the polynomial kernel function. The mentioned accuracy was obtained using 50 population members, and after further analysis, the accuracy value increased to 99.35% with 500 population members.

In order to further develop category models and improve their accuracy, the following can be considered:

- One of the challenges in the field of classification is to develop methods that can separate non-linear structures and in this way perform the classification process with higher accuracy. Non-linear mappings are usually used to increase the flexibility of categories and improve accuracy. In this case, it is suggested to use the polar coordinate axis to improve the performance of cluster models. In such a way that the data are mapped to the polar coordinate axis and then the classification is done on them.
- Since the possible errors in the method used can be different, the use of multiple error rates and cost-sensitive methods in the proposed model of this thesis can lead to its development. In simpler terms, for example, when a patient is mistakenly identified as a healthy person, it creates more cost than the opposite error. Using the multiple error rate can increase the sensitivity of the model to the more expensive error and prevent its occurrence. Considering that the surgical methods of heart patients are invasive methods and the smallest error in the surgery for diagnosis may have irreparable consequences for the health of people, therefore the need for a DSS that can help the doctor in diagnosing the treatment method is very necessary. It seems to ultimately lead to a reduction in treatment costs. Also, DM can help doctors in other diseases such as tuberculosis, heart disease and various cancers.
- The use of data mining in the analysis of medical data is considered a good method to consider existing relationships between variables. It can be seen from the presented method that DM helps to recover useful corrections, even from traits that do not directly indicate the category we intend to predict (heart disease).
- Here the current authors try to predict the probability of heart disease by using the traits that are used to diagnose the disease. It can be expanded to predict other types of diseases that are caused by heart disease. Also, in the future, these data analysis results can be used to increase the accuracy of the prediction system. And to do this, you can use the previous disease diagnosis method for heart patients using the naive method of neural networks with the UCI database. So that we can achieve the best possible accuracy in this method.
- Also, in the future, the combination of ANFIS algorithm and NN can be used in the diagnosis of type 1 heart disease using a standard database such as UCI. And the current authors can also use the fuzzy diagnosis of heart disease based on rules and optimal features based on the combination of DM systems and artificial intelligence algorithms.

REFERENCES

[1] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "**Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm**," Computer methods and programs in biomedicine, vol. 141, pp. 19-26, 2017.

[2] S. Bashir, U. Qamar, F. H. Khan, and L. Naseem, "**HMV: a medical decision support framework using multi-layer classifiers for disease prediction**," Journal of Computational Science, vol. 13, pp. 10-25, 2016.

[3] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "**Heart disease prediction using lazy associative classification**," in 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013, pp. 40-46.

[4] G. Karthiga, C. Preethi, and R. D. H. Devi, "**Heart disease analysis system using data mining techniques**," in 2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14). IEEE, 2014.

[5] T. Amarbayasgalan, K. H. Park, J. Y. Lee, and K. H. Ryu, "**Reconstruction error based deep neural networks for coronary heart disease risk prediction**," PLoS One, vol. 14, p. e0225991, 2019.

[6] P. Gopika, C. Krishnendu, M. H. Chandana, S. Ananthakrishnan, V. Sowmya, E. Gopalakrishnan, et al., "**Single-layer convolution neural network for cardiac disease classification using electrocardiogram signals**," in Deep learning for data analytics, ed: Elsevier, 2020, pp. 21-35.

[7] A. Mehrankia, M. R. Mollakhalili Meybodi, and K. Mirzaie, "**Prediction of Heart Attacks Using Biological Signals Based on Recurrent GMDH Neural Network**," Neural Processing Letters, vol. 54, pp. 987-1008, 2022.

[8] H. Thakkar, V. Shah, H. Yagnik, and M. Shah, "**Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis**," Clinical eHealth, vol. 4, pp. 12-23, 2021.

[9] J. Vijayashree and H. Parveen Sultana, "**Heart disease classification using hybridized Ruzzo-Tompa memetic based deep trained Neocognitron neural network**," Health and Technology, vol. 10, pp. 207-216, 2020.

[10] A. Oztekin, D. Delen, and Z. J. Kong, "**Predicting the graft survival for heart–lung transplantation patients: an integrated data mining methodology**," International journal of medical informatics, vol. 78, pp. e84-e96, 2009.

[11] D. Roa, J. Bautista, N. Rodríguez, M. D. P. Villamil, A. Jiménez, and O. Bernal, "**Data mining: A new opportunity to support the solution of public health issues in Colombia**," in 2011 6th Colombian Computing Congress (CCC), 2011, pp. 1-6.

[12] S. Palaniappan and R. Awang, "**Intelligent heart disease prediction system using data mining techniques**," in 2008 IEEE/ACS international conference on computer systems and applications, 2008, pp. 108-115.

[13] J. R. Quinlan, **C4. 5: programs for machine learning**: Elsevier, 2014.

[14] M. Last and O. Maimon, "**A compact and accurate model for classification**," IEEE Transactions on Knowledge and Data Engineering, vol. 16, pp. 203-215, 2004.

[15] M. Jangizehi, J. Hosseinkhani, A. R. Kenari, and A. Roshandel, "**Providing a Model for Designing a Decision-Making System Using Fuzzy Opinion Mining Process**."

[16] J. Hosseinkhani, S. Ibrahim, S. Chuprat, and J. H. Naniz, "**Web crime mining by means of data mining techniques**," Research Journal of Applied Sciences, Engineering and Technology, vol. 7, pp. 2027-2032, 2014.

[17] K. Leung, K. Lee, J. Wang, E. Y. Ng, H. L. Chan, S. K. Tsui, et al., "**Data mining on dna sequences of hepatitis b virus**," IEEE/ACM transactions on computational biology and bioinformatics, vol. 8, pp. 428-440, 2009.

[18] N. Guru, A. Dahiya, and N. Rajpal, "**Decision support system for heart disease diagnosis using neural network**," Delhi Business Review, vol. 8, pp. 99-101, 2007.

[19] I. Colombet, A. Ruelland, G. Chatellier, F. Gueyffier, P. Degoulet, and M.-C. Jaulent, "**Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression**," in Proceedings of the AMIA Symposium, 2000, p. 156.

[20] A. KR and A. UM, "**Ethical and legal issues for medical data mining**," international Journal of Computer Applications, vol. 1, p. 7, 2010.

[21] M. Kantardzic, **Data mining: concepts, models, methods, and algorithms**: John Wiley & Sons, 2011.

[22] G. Chen and T. Åstebro, "**How to deal with missing categorical data: Test of a simple Bayesian method**," Organizational Research Methods, vol. 6, pp. 309-327, 2003.

[23] M. Anbarasi, E. Anupriya, and N. Iyengar, "**Enhanced prediction of heart disease with feature subset selection using genetic algorithm**," International Journal of Engineering Science and Technology, vol. 2, pp. 5370-5376, 2010.

[24] J. Hosseinkhani, H. Taherdoost, and S. Keikhaee, "**ANTON framework based on semantic focused crawler to support web crime mining using SVM**," Annals of Data Science, vol. 8, pp. 227-240, 2021.