Transactions on Machine Intelligence 6(2) (2023) 114-125

ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir/index.php/TMI

Cardiovascular Disease Diagnosis Using the Combination of Principal Component Analysis Algorithm and Regression Tree

H. R. Aviny[1,*], M. Ghasemi[2], M. Fazlazad[3]

[1]Ph.D. student, Department of Computer Engineering, Faculty of Technology and Engineering, Yasouj branch, Islamic Azad University, Yasouj, Iran.
[2]Masters' student, Department of Computer Engineering, Faculty of Technology and Engineering, Yasouj branch, Islamic Azad University, Yasouj, Iran.
[3]Masters, Department of Computer Engineering, Faculty of Technology and Engineering, Yasouj branch, Islamic Azad University, Yasouj, Iran.

ARTICLE INFO

Article History:
Received 10 April 2023
Received in revised form 20 May 2023
Accepted 26 June 2023
Available online 30 June 2023

Keywords:
Cardiovascular Disease, Principal Component Analysis Algorithm, Regression Tree Algorithm

ABSTRACT

Cardiovascular disease stands as a prominent global cause of mortality, emphasizing the pivotal need for effective diagnostic and treatment strategies. Recognizing the significance of early detection, this study centers on employing the regression tree algorithm as a primary method. To gauge the precision of cardiovascular disease diagnosis, we scrutinized a dataset encompassing 270 patient samples and 14 distinct characteristics. The implementation approach involved a dual deployment of the Principal Component Analysis (PCA) algorithm and the regression tree algorithm. Employing PCA, we streamlined the feature set from 14 to 8, followed by the application of the regression tree algorithm to enhance detection accuracy. The decision tree classification method adopted encompasses critical facets such as feature selection, tree generation, and pruning. Implementation of these procedures was facilitated through the Weka tool, a data mining software. The collaborative utilization of PCA and the regression tree algorithm culminated in a noteworthy improvement, yielding a diagnostic accuracy increase of 81.48% in detecting cardiovascular disease.

1. INTRODUCTION

Throughout history, illnesses have posed the greatest danger to humanity. However, heart disease has received more attention in medical research. In recent years, there has been extensive investigation into the classification and diagnosis of heart disease as a crucial topic. Numerous studies have been undertaken to enhance accuracy and minimize errors in making these diagnoses [1]. With the development of AI methods, such as ML and DL, the models based on these methods will soon be an inseparable part of diagnostic equipment in the field of coronary

* Corresponding Author: hamidreza.avini93@gmail.com
Department of Computer Engineering, Faculty of Technology and Engineering, Yasouj branch, Islamic Azad University, Yasouj, Iran

http://dx.doi.org/10.47176/TMI.2023.114

artery disease. The employment of these tools paves the way for providing clinical specialists with specialized consultations in the CAD detection area. As instruments in clinical specialists' hands, these models as screening software modules prevent risky and invasive diagnostic tests and take the high financial burden of CAD detection and other coronary artery diseases from the shoulders of clinical care systems [2]. Proper heart function is essential for human life, as impaired function can negatively impact other bodily systems, including the mind and kidneys. Cardiovascular diseases are the leading cause of death globally [3], and as such, research into new bio spraying methods has garnered significant attention [4]. Cardiovascular diseases, also known as heart disease, are a group of problems primarily connected to a process called atherosclerosis. Atherosclerosis happens when fatty deposits, or plaque, accumulate in the walls of arteries. This plaque buildup restricts blood vessels, making it harder for blood to flow through them. In some instances, a blood clot may form at the narrowed point, causing blood flow to be completely cut off. Such a condition is called a stroke, which can occur in the brain or heart [5].

According to the World Health Organization, heart diseases cause 12 million deaths worldwide. Cardiovascular disease is the leading cause of death in Iran, with 38% of Iranians succumbing to the illness [3]. Heart attack is the primary cause of death for those aged over 35 in Iran [4]. At the start of the 20th century, cardiovascular diseases were responsible for 10% of global deaths. By the end of the 20th century, fatalities resulting from heart disease rose to 25%. Based on current trends, it is estimated that cardiovascular disease will be responsible for over 35% to 60% of all global deaths by 2025. In Iran, 44% of deaths are attributed to cardiovascular diseases. Thus, considering the widespread prevalence of heart diseases, it is imperative to employ data mining methods for the diagnosis of cardiovascular diseases [6].

We utilized a combination of a dimensionality reduction algorithm and tree regression to enhance the accuracy of diagnosing cardiovascular disease. The implementation was performed via WEKA software. Our approach resulted in a significant boost in accuracy for cardiovascular disease diagnosis. The article is structured as follows: The second section provides background information on the research, the third section presents the dimension reduction algorithm, the fourth section outlines the tree regression algorithm, the fifth section covers the research findings, and finally, the sixth section presents the conclusion.

Metabolic syndrome (MetS) has been found to increase the risk of cardiovascular-related adverse events among patients with CVD [7]. Furthermore, the study by Li et al. (2021) reported that the proposed MLP-PSO outperforms all other algorithms, obtaining an accuracy of 84.61% for predicting heart disease [8]. This suggests that machine learning techniques can be effective in identifying CVD and its related risk factors. The important relationship between obesity, insulin resistance, and oxidative stress in individuals with MetS has been highlighted by Muthiah et al. (2021) and Jakubiak et al. (2021) [9,10]. These components of MetS are crucial in understanding the pathophysiology of CVD and could potentially be used as input features for machine learning models to enhance CVD diagnosis.

In the study by Zhu et al. (2021), the random tree model performed admirably, achieving the highest accuracy of 100% for predicting CVD patients. This indicates the potential of using regression tree models in combination with PCA for accurate CVD diagnosis [11]. Moreover, Nadakinamani et al. (2022) demonstrated that clinical data analysis using machine learning techniques can aid in predicting CVD, further emphasizing the utility of such methods [12]. The association between COVID-19 and short- and long-term risk of CVD and mortality was investigated by Wan et al. (2023) [7]. This highlights the need for accurate and efficient diagnostic tools to identify individuals at risk of developing CVD, especially in the context of emerging health crises such as the COVID-19 pandemic.

Although machine learning algorithms such as MLP-PSO and random tree models have shown promising results in CVD diagnosis, there are still knowledge gaps that require further exploration [13]. Future research should focus on validating these findings through large-scale clinical trials and real-world applications [14]. Additionally, investigating the incorporation of epigenetic regulation and other clinical biomarkers in machine learning models for CVD diagnosis could enhance the accuracy and reliability of these algorithms. Lastly, exploring the potential impact of environmental and lifestyle factors on CVD risk prediction using machine learning techniques would provide a more comprehensive understanding of CVD diagnosis and prevention.

In summary, the combination of PCA algorithm and regression tree models, along with machine learning techniques, holds great promise for improving the diagnosis of CVD. The findings from the existing literature suggest the need for further research to fully realize the potential of these models in clinical practice and public health interventions.

2. CARDIOVASCULAR DISEASE

Heart disease is a broad term for conditions that affect the structure and function of the heart muscle. Cardiovascular diseases are a group of diseases that include coronary heart disease, arrhythmia, heart failure and valve disease [15].

2.1. All types of cardiovascular diseases

- Irregular heartbeat (arrhythmia)
- Congenital heart defect
- Weak heart muscles (cardiomyopathy)
- Heart valve problems
- Heart infection
- Cardiovascular disease

2.2. The most common types of heart disease

The most common heart diseases are:

- Heart failure
- Congenital heart defects
- Peripheral vascular disease
- Among other diseases of the heart and blood vessels are heart valve diseases
- Uncontrollable risk factors in cardiovascular diseases
- Controllable risk factors in cardiovascular diseases
- Silent ischemia
- Arrhythmia
- Heart defect
- Angina
- Hereditary heart diseases

2.2.1. Heart failure

In this disease, the ability of the heart to pump blood becomes weaker than normal. With the weakness of the heart in pumping blood, the movement of blood through the heart and body is done at a slower speed. This slowness in blood movement in turn forces the heart to do more work to deliver blood and oxygen to the organs, on the other hand, the extra work of the heart causes fatigue and shortness of breath for the affected person. Some symptoms of heart failure [15]:

1- Feeling of shortness of breath when lying down too much, 2- Fatigue and weakness, 3- Swelling in the legs and ankles, 4- Continuous cough or wheezing with white or pink sputum stained with blood, 5- Increased need to urinate at night, 6- Abdominal swelling, anorexia and nausea, 7- Sudden weight gains due to fluid retention

2.2.2. Congenital heart defects

One of the types of heart diseases is defects that exist in the heart at birth. These defects are not considered diseases, but they are called disorders. Disorders caused in the heart occur during the formation and growth of the fetus in the mother's body [15]. Symptoms of congenital heart defects:

1- Heart palpitations (arrhythmia), 2- Bluish colour of the skin (cyanosis), 3- Shortness of breath, 4- Getting tired quickly due to exercise, 5- Dizziness or fainting, 6- Swelling of body tissue or organs

Other possible symptoms related to congenital heart disease are:

1- Excessive sweating, 2- Feeling very tired and bruised, 3- Rapid heartbeat, 4- Chest pain, 5- Blue colour or bruise on the skin, 6- Sticking of nails

2.2.2.1. Effects of congenital heart disease

Congenital heart diseases, based on their severity, include a wide range, ranging from simple complications such as the presence of holes between the chambers of the heart to very severe abnormalities such as the complete absence of one or more heart chambers or valves. In addition, having a congenital heart disease can increase the risk of some other diseases, including the following [15]:

1- Increased pulmonary blood pressure, 2- Cardiac arrhythmia or irregular heartbeat, 3- Internal infection of the heart, 4- Absence of blood coagulation, 5- Congestive heart failure

2.2.2.2. Prenatal care to prevent congenital heart diseases

Because the main cause of most congenital heart diseases is not precisely known, it is not possible to prevent these diseases. However, there are several care measures that can be followed to reduce the overall risk of birth defects, including heart disease, in children. Some of these measures include [15]:

- Rubella vaccination for rubella infection during pregnancy can lead to heart diseases in the baby. So be sure to get the rubella vaccine before you get pregnant [15].
- Avoid harmful substances. During pregnancy, avoid painting the building or washing with detergents that have a strong smell. Also, avoid taking medicines, herbal substances and food supplements without consulting your doctor. Do not smoke or drink alcohol during pregnancy [15].
- Take multivitamins and folic acid. According to research, daily consumption of 400 micrograms of folic acid reduces the possibility of congenital malformations in the brain and spinal cord of the child, as well as reducing the risk of congenital heart diseases [15].

2.2.3. Peripheral vascular disease

One of the types of diseases that affect the circulatory system is peripheral vascular disease. This heart disease occurs when fat and cholesterol deposits, or plaque, form in the peripheral arteries, which are the blood vessels outside the heart. The creation of these plaques is called atherosclerosis. Arterial narrowing means limiting the amount of blood flow to body tissues. Depending on the artery in which the blockage occurs, this blockage can lead to stroke, heart attack, kidney disease and other serious diseases. The symptoms of this disease include [15]:

1- Painful cramping in the hip joint or leg muscles after certain activities, such as walking or climbing stairs, 2- Leg numbness or weakness, 3- Coldness in the leg, especially when compared to another part of the body, 4- Late healing of wounds on the toes, 5- Change in the color of the legs, 6- Hair loss or slower growth of hair in the legs, 7- Slower toenail growth, 8- Shining of the skin of the feet, 9- Weak pulse in the legs, 10- Erectile dysfunction in men

2.2.4. Among other diseases of the heart and blood vessels are heart valve diseases

In the human heart, there are several valves whose task is to direct the blood in a specific direction. If these valves suffer from functional failure or narrowing and obstruction, they significantly disrupt blood flow in the heart. Mitral valve dilatation, aortic valve obstruction and stenosis, mitral valve dysfunction, pulmonary valve obstruction and stenosis are among the most common types of heart valve diseases [15].

2.2.5. Uncontrollable risk factors in cardiovascular diseases

1- Age: The risk of heart disease increases with age, 2- Gender: Men are affected by these diseases more than women. But women get sick faster, 3- Family history

2.2.6. Controllable risk factors in cardiovascular diseases

1- Cholesterol, 2- Level of Physical Activity, 3- Smoking, 4- Weight, 5- Blood Pressure, 6- Diabetes

2.2.7. Silent ischemia

This disease is a type of coronary artery disease in which blood flow to the heart muscle is reduced but causes much less pain or symptoms. When discomfort is felt, physical activity usually occurs [15].

2.2.8. Arrhythmia

Cardiac arrhythmia means irregular heartbeat, arrhythmia means a problem in the heartbeat rhythm. This disease occurs when the electrical impulses that regulate the heartbeat are disrupted and do not work properly. Types of rhythm or irregular heartbeat include [15]:

1- Tachycardia: It means when the heart beats very fast, 2- Bradycardia: It means when the heart beats very slowly, 3- Fibrillation: It is when the heartbeat is irregular.

Irregular heartbeat is common and most of us experience it, but if this heartbeat is very different from regular heartbeat or due to injury or weakness of the heart, it should be taken seriously. An irregular heartbeat can be fatal. When this irregular heartbeat is prolonged or causes chest pain, you should go to the hospital immediately [15].

2.2.9. Heart Defect

An occlusion is a heart defect that partially or completely blocks blood flow. A narrowing can occur in the heart valves, arteries, or veins, which is called an occlusion [15].

1-Aortic stenosis 2- Aortic valve 3- Mitral valve prolapses 4- Pulmonary stenosis 5- Subaortic stenosis

2.2.10. Angina

Angina, which is also known as angina pectoris, is a disease in which enough oxygen does not reach the heart. Although angina is not technically a disease, it is one of the symptoms of coronary artery disease, because the lack of oxygen is caused by the closing of the coronary arteries [15].

2.2.11. Hereditary heart diseases

Genes affect the appearance and functioning of the body and make each person unique. Many heart conditions are also hereditary, so if one of a person's parents has a problematic gene, there is a 50% chance that you will also have that gene [15]. This hereditary condition is caused by the defect or mutation of one or more genes in the body, and if not treated, it may be life-threatening. Hereditary heart diseases affect people at any age. It is especially difficult to bring up this disease with the family, because they may also carry the same problematic genes, even if the problematic gene is inherited, it is impossible to say how it affects people [15].

2.3. Drug treatment of cardiovascular disease

This treatment includes the use of drugs that help reduce the severity of the disease and its damage, these drugs include [15]:

1- Antihypertensive drugs, 2- Medicines that reduce heart palpitations, 3- Cholesterol-lowering drugs, 4- Medicines to stabilize heart rate, 5- Drugs to prevent blood clots in the coronary arteries, 6-Medicines to improve blood pumping in the body of a person with heart disease

Other ways to treat heart disease include angioplasty, bypass surgery, or other surgeries.

2.4. Prevention of heart diseases

By changing the lifestyle to a healthy one, the risk of heart diseases can be prevented. The basic factors of a healthy lifestyle include [15]:

1- Not smoking or quitting smoking, 2- At least 30 minutes of exercise a day, 3- Proper diet (lots of vegetables and fruits and little fat, sugar and meat), 4- Avoid alcohol consumption, 5- Controlling diseases such as diabetes, high blood pressure and cholesterol, 6- Asking your family to help you with the above changes, although heart disease is treatable, preventing heart disease with lifestyle changes seems more logical than any other action.

3. RESEARCH BACKGROUND

Various data mining methods have been utilized to diagnose cardiovascular disease. Below, we describe some of the most vital techniques employed for diagnosis.

3.1. Cardiovascular disease diagnosis using regression classification algorithm

In this article, for the diagnosis of cardiovascular diseases, a dataset including 270 samples and 14 characteristics have been examined. Regression classification algorithm is used for implementation. The accuracy of detection obtained with this algorithm is 80% [16].

3.2. Prediction of heart disease treatment method using data mining algorithms

In this article, the C&R Tree algorithm has been used for the accuracy of cardiovascular disease. This algorithm is a decision tree that has two capabilities of classification and regression. The accuracy of cardiovascular disease diagnosis using this algorithm is 76.04%. [3].

4. PRINCIPAL COMPONENT ANALYSIS ALGORITHM

Dimensionality reduction is a technique employed to highlight variation and establish robust patterns within data sets. The approach can effectively identify principal components and facilitate the analysis of critical features rather than examining the entire set. Specifically, reducing the dimensionality extracts the most valuable features for analysis [17].

4.1. Analysis of the main component of Principal Component Analysis algorithm

This approach is a primary linear method for reducing dimensions [17-18]. It involves linearly mapping data to a lower-dimensional space that explains the variance of the initial data in the transferred data. To achieve this reduction, the variance (and occasionally covariance) matrix of the data is formed, and the eigenvector of that matrix is computed. The principal vectors associated with the highest eigenvalues hold the most informative content of the original dataset and can be leveraged to regenerate a substantial part of its variance. However, efforts are made to

ensure that the variance remains explanatory to a considerable degree (with the expectation of sustaining the quality of the findings) [17]. Initially, the first few vectors represent the performance of the major data. Employing these critical vectors, the data in lower dimensions is formulated with some loss of information.

4.2. feature extraction

Data in high-dimensional spaces can be transformed into lower-dimensional spaces through feature extraction, which can be achieved through linear methods like principal component analysis or non-linear methods. Non-linear techniques are often faster and simpler, but when dealing with complex data, non-linear methods are typically more accurate. Multidimensional data can be processed using the tensor representation and multilinear subspace learning for dimensionality reduction [17].

4.3. The core of dimension reduction

Principal component analysis can be performed non-linearly using the kernel method. The resulting technique is able to construct non-linear mappings that maximize the amount of variance explained in data in a smaller space. In general, this method uses a method similar to PCA, with the difference that instead of using a linear mapping from the data in low dimensions to the data in the main dimensions using the kernel method, it uses non-linear mapping [17].

5. TREE REGRESSION CLASSIFICATION ALGORITHM (CART)

Regression tree classification algorithm was proposed by Leo Berryman, Jerome Friedman, Richard Olsen and Charles Stone in 1984. Its application is very wide because it is one of the important classification methods of decision tree. Feature selection, tree generation and pruning are important parts of tree regression. The segmentation method using binary regression is used in tree regression. The conditional probability distribution of variable Y is based on the input random variable X output, and the current set is divided into two subsets, making it a sub-node of the decision tree [19].

The CART decision tree utilizes the gini index to make node selections. Conversely, ID3 and C4.5 trees utilized entropy and gain. To determine features that provide more information, the CART tree relies on the Gini index. In essence, the lower the Gini index for each feature (dimension), the more information the feature provides. If a feature in a dimension has a lower Gini index, it contains more information and should be placed closer to the root in the constructed tree. The tree uses a trial-and-error approach to determine the optimal value for the direction of the separator point in each feature dimension. The distinction between the Gini index and entropy lies in the fact that the former is typically effective for data with a larger portion, while the latter is useful for data with numerous small parts containing unique values [19].

Decision trees, including the CART decision tree, are susceptible to overfitting. To prevent overfitting in the CART decision tree, a stopping condition can be implemented, instructing the algorithm to halt the tree's progression. This technique ensures that the tree stops branching and creating new leaves when the sample size in a sub-tree falls below the threshold value [19]. Using a specific number of samples in a sub-tree is one method to prevent the CART tree from growing beyond a certain threshold, as complexity can lead to overfitting in classification models.

The regression tree classification process is a process in which the training set is divided into smaller and smaller subsets. The ideal result is that the same label exists for the leaf samples to generate the tree. The selection criterion of tree regression nodes is to minimize the impurity of nodes as much as possible. The lowest Gini coefficient for each feature is used as a standard for selecting test features in tree regression. Suppose there are k classes, the probability that a sample belongs to class k is in Pk, then there is a Gini coefficient for the probability distribution [19]:

$$Gini = 1 - \sum_{n=1}^{k} P_K^2 \quad (1)$$

If the eigenvalue of the set T is A, then the set T can be divided into two parts of T1 and T2 according to the eigenvalues. Then the Gini coefficient of the set T is defined as below [19]:

$$Gini(T,A) = \frac{|T_1|}{|T|} Gini(T_1) + \frac{|T_2|}{|T|} Gini(T_2) \quad (2)$$

The larger the Gini coefficient calculated from Gini(T,A), the greater the uncertainty of the sample set. The accuracy of the decision tree created through tree regression algorithm is high, while the complexity is not very high [19].

5.1. Description of tree regression classification algorithm

Tree regression classification algorithm is a non-parametric model and without any defaults, it is used to measure the relationship between independent variables and dependent variable or target, and it is one of the important methods of data mining, it has been widely used in business, industry, engineering and other sciences. Tree regression is a powerful tool in determining the most important independent variables and solving classification and prediction problems. In this algorithm, there are a number of records whose categories are already known. The goal is to prepare a tree by which the dependent variable or the same class can be predicted and determined for a new record. The tree regression method creates its branches in pairs and only based on one field (independent variable). That is, every group other than its leaves is divided into two other groups. By non-leaf group, we mean the node of the tree model, which itself is separated into two other parts.

The first step is to answer the question of which field produces the best branch. The best branching occurs when the resulting branches are such that in each branch one class dominates the other classes. The standard that is used to evaluate the branches is diversity. There are many methods to calculate diversity in a set of records, in all of them, high diversity is the set that has different classes in it, and low diversity is the set. The ones where the members of one class overpower other classes and the best way to create a branch is to reduce diversity in collections as much as possible. In the next step, there are two branches, each of which has a series of records, each of the records of the higher node is placed in one of the branches. That is, for each of them, the field is selected again so that the best new branches can be created with minimum variety. This stage continues until a node is produced in each sub-branch so that the creation of a new branch in that node does not significantly reduce the amount of diversity. This final node is called a leaf. To separate each node into two sub-nodes, there are various indices, the most famous of which for nominal data is the Gini index, which is defined in the form of equation (3) [19]:

$$p(j|m) = \frac{p(j,m)}{p(m)}, p(j,m) = \frac{\pi(j) n_j(m)}{n_j}, p(m) = \sum_{j=1}^{J} p(j,m) \quad (3)$$

$$Gini(m) = 1 - \sum_{J=1}^{J} P^2(j|m) \quad (4)$$

In elucidating the intricacies of the decision tree model, let's employ a systematic breakdown of the key components involved. The variable *J* represents the count of target variables or categories. The parameter *p(j)* signifies the initial probability linked to category *j*. Delving into the nodes, *n_j(m)* denotes the number of observations specific to category *j* in node m, while *n_j* accounts for the overall count of observations affiliated with category j in the root node.

Moving on, *p((j)|m)* represents the probability of placing observations related to category j in node m, and Gini(m) stands as the Gini index, serving as an indicator of impurity with heterogeneity at node m. A Gini(m) value of zero within a node implies a scenario where all observations in that node belong to the same category, denoting maximum purity. Conversely, the highest Gini(m) value is reached when all observations within the node are proportionally identical. The iterative Gini index computation spans all nodes and variables. The separator variable, pivotal for decision tree branching, is determined based on the variable with the lowest Gini value. The initial probability *(p(j))* sheds light on each category's contribution to the overall reference.

The growth of the tree, facilitated by the Gini index, initiates at the first node, encompassing all observations. In the creation of each tree, equation (5) is employed to calculate the cost of misclassification, serving as a robust fit index [19]. This meticulous approach ensures the development of a decision tree model that effectively navigates the complexities of multi-category classification with precision and reliability.

$$misclassificationcost = \sum_{t=1}^{T} p(t)\left[1 - \sum_{j=1}^{J} p^2(j|t)\right] \quad (5)$$

In this context, let's delve into the intricacies of the decision tree model. Let *J* symbolize the number of target variables or categories, with *p(j)* denoting the initial probability associated with category *j*. Moving to the nodes, *n_j(m)* signifies the number of observations linked to category j in node m, while *n_j* represents the overall count of observations for category j in the root node.

To further elucidate, *p((j)|m)* represents the probability of placing observations related to category *j* in node m, and Gini(m) encapsulates the Gini index, providing insights into the purity level at node m. A Gini index of zero within a node implies all observations in that node belong to the same category, signifying maximum purity. Conversely, the maximum Gini(m) value occurs when the observations within the node are proportionally identical.

The iterative process of Gini index calculation unfolds across all nodes and variables. The separator variable, crucial for decision tree branching, is selected based on the variable with the lowest Gini value. The initial probability offers insight into each category's contribution to the overall reference. Commencing at the first node and encompassing all observations, the tree's growth utilizes the Gini index. Equation (5) is then employed for each tree to compute the cost of misclassification, serving as an excellent fit index [19]. This comprehensive approach ensures a robust decision tree model that efficiently navigates the complex landscape of multi-category classification.

6. RESEARCH FINDINGS

This section will briefly explain the implementation method and evaluation criteria.

6.1. Implementation method

For the implementation, we use the Weka tool, which is a data mining tool, and the implementation method includes 5 steps: 1. Data preparation step 2. Dimension reduction step 3. Sampling step 4- Modeling 5- Data classification step the steps are fully explained. Data set: A data set containing 270 samples (patients) was uploaded from the UCI website, in which 14 features were analyzed in each sample.

Table 1. Characteristics of cardiovascular disease diagnosis [16].

Property	Description
Age	Age
Sex	value 0 "sir" value 1 "ma'am"
CP	Type of chest pain (= "1" typical angina, = "2" atypical angina, = "3" non-angina pain = "4" asymptomatic)
Trustbps	Blood pressure at rest in mm Hg
Chol	Serum cholesterol in mg/dl.
Fbs	Fasting blood sugar has a value of "1" if it is greater than 120 and "0" if it is less than 120.
Restecg	ECG results at rest = "0" (normal, ="1" abnormal ST-T wave, ="2" possibility or certainty of left ventricular hypertrophy)
Thalach	Maximum heart rate achieved.
Ex	The indicator that angina is caused by exercise (= "1" yes, "0" no).
Oldpeak	Exercise-induced ST reduction compared to others.
Slope	The peak of the sports slope of the ST department
ca	Number of main vessels colored by fluoroscopy.
Thal	Summary of heart disease ("3 = normal, "6 = fixed defect, "7 = reversible defect).
Num	Values (1,2,3,4) and no disease (and value 0)

6.1.1. Data preparation

In this step, the dataset is added to the Weka tool.

6.1.2. Dimension reduction algorithm

By using dimension reduction, we reduce the number of features from 14 to 8 as shown in table (2).

Table 2. Reducing the dimensions of the number of features from 14 to 8.

Property	Description
Age	Age
Sex	value 0 "sir" value 1 "ma'am"
CP	Type of chest pain (= "1" typical angina, = "2" atypical angina, = "3" non-angina pain = "4" asymptomatic)
Trustbps	Blood pressure at rest in mm Hg
Chol	Serum cholesterol in mg/dl.
Fbs	Fasting blood sugar has a value of "1" if it is greater than 120 and "0" if it is less than 120.
Restecg	ECG results at rest = "0" (normal, ="1" abnormal ST-T wave, ="2" possibility or certainty of left ventricular hypertrophy)
Thalach	Maximum heart rate achieved.

6.1.3. Data sampling

In this step, we select the number of training and test datasets, comprising seed=1 and fold=10.

6.1.4. Tree regression algorithm

This step--which was explained in the previous section--involves adding the algorithm to the tool to implement control.

6.1.5. Classification of data

Based on the proportion of classified test samples or data sets, an estimation of the classification accuracy is provided. Table 3 displays the implementation outcomes.

Table 3. Results of implementing the combination of Principal Component Analysis algorithm + regression tree algorithm

Number	Evaluation criteria	PCA+CART
1	Correctly The number of correctly classified samples	48.81% 220
2	Incorrectly Number of misclassified samples	52.18% 50
3	Kappa	0.61%
4	Mean absolute error	0.26%
5	Root mean squared error	0.38
6	Relative absolute error	53.26%
7	Root relative squared error	78.17%
8	Total Number of Instances	270

6.2. Evaluation criteria

Based on Table 3, there are 220 accurately classified samples, leading to an accuracy of 81.48% using Equation (6) to calculate.

$$Accuracy = \frac{TP^{\dagger}+TN^{\ddagger}}{TP+TN+FP^{\S}+FN^{**}} \quad (6)$$

Based on Table 3, there were 50 incorrectly classified samples, resulting in a classification error rate of 18.52%. Equation (7) was used to calculate the amount of error.

$$Error = 100 - \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

7. CONCLUSION

Cardiovascular disease, a pervasive global cause of mortality, necessitates heightened diagnostic accuracy. This imperative led to the synergistic utilization of the dimensionality reduction algorithm and the tree regression algorithm. The latter, a pivotal component in decision tree classification, encapsulates essential elements such as feature selection, tree generation, and pruning. The amalgamation of these algorithms culminated in a notable 81.48% increase in accuracy for cardiovascular disease diagnosis.

Furthermore, it is plausible that the landscape of diagnostic accuracy could witness further refinement through the exploration of diverse data mining algorithms in the future. The evolving field of data analytics holds promise for the continual improvement of cardiovascular disease diagnosis. The ongoing exploration and integration of advanced algorithms stand as a beacon, guiding us toward enhanced precision and efficacy in the critical realm of cardiovascular health assessment. Embracing the potential of innovative approaches ensures that we are well-positioned to make strides in the ongoing battle against this leading global cause of mortality.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Harvi, M., Satayshi, S. (2013), the title of the article Smart and fast diagnosis of heart disease based on synergy of linear neural networks and logical regression method, place of publication: Journal of Mazandaran University of Medical Sciences, volume 24, number 12, page 11.

[2] Garavand, A., Behmanesh, A., Aslani, N., Sadeghsalehi, H., & Ghaderzadeh, M. (2023). Towards diagnostic aided systems in coronary artery disease detection: a comprehensive multiview survey of the state of the art. International Journal of Intelligent Systems, 2023.

[3] Mazaheri, S., Ashuri, M., Bechari, Z. (2016), the title of the article predicting the method of treatment of heart disease using data mining algorithms, place of publication: paramedical journal of Tehran Faculty of Medical Sciences, volume 11, number 3, page 10.

[4] Safdari, R., Ghazi Saeedi, M., Qaroni, M., Nasbari, M., & Erji, G. (2014). comparing the performance of decision tree and neural network in predicting cardiac infarction. Mashhad Journal of Paramedical Sciences and Rehabilitation, 3(2).

[5] Abidi Pharmaceutical Co. (n.d.). *Common cardiovascular diseases*. Abidi Pharma. Retrieved May 30, 2023, from https://abidipharma.com/health-items/common-cardiovascular-diseases/

[6] Noor, J., Saadi, Z., & Akbari, A. (2017). the title of the article is designing a decision support system to diagnose and predict heart disease using artificial neural network; A case study of Ayatollah Golpayegani

† - True Positive
‡ - True negative
§ - False Positive
** - False Negative

Hospital in Qom, publication: Management strategies in the health system.

[7] Wan, E.., Mathur, S.., Zhang, Ran., Yan, Vincent K. C.., Lai, F.., Chui, C.., Li, Xia., Wong, C.., Chan, E.., Yiu, K.., & Wong, I.. (2023). Association of COVID-19 with short- and long-term risk of cardiovascular disease and mortality: a prospective cohort in UK Biobank.. Cardiovascular research . http://doi.org/10.1093/cvr/cvac195

[8] Li, Xiao., Zhai, Ya-jing., Zhao, Jiaguo., He, Hairong., Li, Yuan-jie., Liu, Yue., Feng, Aozi., Li, Li., Huang, Tao., Xu, Anding., & Lyu, Jun. (2021). Impact of Metabolic Syndrome and It's Components on Prognosis in Patients With Cardiovascular Diseases: A Meta-Analysis. Frontiers in Cardiovascular Medicine, 8. http://doi.org/10.3389/fcvm.2021.704145

[9] Muthiah, M.., Han, Ng Cheng., & Sanyal, A.. (2021). A clinical overview of non-alcoholic fatty liver disease: A guide to diagnosis, the clinical features, and complications—What the non-specialist needs to know. Diabetes , 24 , 14 - 3 . http://doi.org/10.1111/dom.14521

[10] Jakubiak, G.., Osadnik, K.., Lejawa, M.., Osadnik, T.., Goławski, M.., Lewandowski, P.., & Pawlas, N.. (2021). “Obesity and Insulin Resistance” Is the Component of the Metabolic Syndrome Most Strongly Associated with Oxidative Stress. Antioxidants , 11 . http://doi.org/10.3390/antiox11010079

[11] Zhu, Rong., Wang, Yong., Liu, Jin-Xing., & Dai, Lingyun. (2021). IPCARF: improving lncRNA-disease association prediction using incremental principal component analysis feature selection and a random forest classifier. BMC Bioinformatics , 22 . http://doi.org/10.1186/s12859-021-04104-9

[12] Nadakinamani, RajkumarGangappa., Reyana, A.., Kautish, Sandeep., Vibith, A. S.., Gupta, Yogita., Abdelwahab, S.., & Mohamed, A. W.. (2022). Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques. Computational Intelligence and Neuroscience , 2022. http://doi.org/10.1155/2022/2973324

[13] Bataineh, Ali Al., & Manacek, Sarah. (2022). MLP-PSO Hybrid Algorithm for Heart Disease Prediction. Journal of Personalized Medicine , 12 . http://doi.org/10.3390/jpm12081208

[14] Shi, Yuncong., Zhang, Huanji., Huang, Suli., Yin, Li., Wang, Feng., Luo, Pei., & Huang, Hui. (2022). Epigenetic regulation in cardiovascular disease: mechanisms and advances in clinical trials. Signal Transduction and Targeted Therapy , 7 . http://doi.org/10.1038/s41392-022-01055-2

[15] Afshari, N. (n.d.). *Frequently Asked Questions – Page 195*. Dr. Nader Afshari. Retrieved May 30, 2023, from https://drnaderafshari.ir/faq/?cat=all&hpage=195

[16] Aviny, M., & Armin, M. (Eds.). (2019). Cardiovascular Disease Diagnosis Using Regression Classification Algorithm. In 5th Iran National Conference on Computer Engineering and Blockchain.

[17] Safdari, R., Ghazi Saeedi, M., Qaroni, M., Nasbari, M., & Erji, G. (2014). comparing the performance of decision tree and neural network in predicting cardiac infarction. Mashhad Journal of Paramedical Sciences and Rehabilitation, 3(2).

[18] Noor, J., Saadi, Z., & Akbari, A. (2017). the title of the article is designing a decision support system to diagnose and predict heart disease using artificial neural network; A case study of Ayatollah Golpayegani Hospital in Qom, publication: Management strategies in the health system.

[19] Gholami, F. (2017). diagnosis of cardiovascular disease using the combination of dimensionality reduction algorithm and simple Bayesian tree algorithm, the third national conference in computer engineering, information technology and data processing. 9.