I.J. Information Technology and Computer Science, **2022, 6, 23-32**
Published Online on December 8, 2022 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2022.06.03

Comparative Analysis of Data Mining Techniques to Predict Cardiovascular Disease

Md. Al Muzahid Nayim*
Department of Computer Science, Faculty of science and technology, American International University-Bangladesh (AIUB), Dhaka, Bangladesh
E-mail: almuzahidniem@gmail.com
ORCID iD: https://orcid.org/0000-0002-0651-2221
*Corresponding Author

Fahmidul Alam
Department of Computer Science, Faculty of science and technology, American International University-Bangladesh (AIUB), Dhaka, Bangladesh
E-mail: fahmidul.alam.bd@gmail.com
ORCID iD: https://orcid.org/0000-0002-2262-3597

Md. Rasel
Department of Computer Science, Faculty of science and technology, American International University-Bangladesh (AIUB), Dhaka, Bangladesh
E-mail: rasel.aiub990@gmail.com
ORCID iD: https://orcid.org/0000-0002-2928-6984

Ragib Shahriar
Department of Computer Science, Faculty of science and technology, American International University-Bangladesh (AIUB), Dhaka, Bangladesh
E-mail: ragib.hridoy@gmail.com
ORCID iD: https://orcid.org/0000-0001-7699-0969

Dip Nandi
Faculty of Science and Technology, American International University-Bangladesh (AIUB), Dhaka, Bangladesh
E-mail: dip.nandi@aiub.edu
ORCID iD: https://orcid.org/0000-0002-9019-9740

Received: 10 August 2022; Revised: 26 September 2022; Accepted: 14 October 2022; Published: 08 December 2022

Abstract: Cardiovascular disease is the leading cause of death. In recent days, most people are living with cardiovascular disease because of their unhealthy lifestyle and the most alarming issue is the majority of them do not get any symptoms in the early stage. This is why this disease is becoming more deadly. However, medical science has a large amount of data regarding cardiovascular disease, so this data can be used to apply data mining techniques to predict cardiovascular disease at the early stage to reduce its deadly effect. Here, five data mining classification techniques, such as: Naïve Bayes, K-Nearest Neighbors, Support Vector Machine, Random Forest and Decision Tree were implemented in the WEKA tool to get the best accuracy rate and a dataset of 12 attributes with more than 300 instances was used to apply all the data mining techniques to get the best accuracy rate. After doing this research people who are at the early stage of cardiovascular disease or probably going to be a victim can be identified more accurately.

Index Terms: Data Mining, WEKA, Classification Techniques, Cardiovascular Disease (CVD).

1. Introduction

Cardiovascular disease (CVD) has become world's one in all the foremost common diseases in last decades. In recent days, most people are living with cardiovascular disease because of their unhealthy lifestyle and the most

alarming issue is 50% of them do not get any symptoms in the early stage [1]. 90% of total death rate for CVD is account for the late diagnosed [2]. At the end of this research, the results would be helpful to reduce this problem by increasing the accuracy rate of CVD investigation techniques. In medical science, there are some popular investigation techniques that are used to identify CVD. But these investigation techniques also have some limitations.

This research document is aimed to identify cardiovascular disease using data mining techniques at an early age so that the death toll of cardiovascular disease can be decreased. Medical research now has a vast quantity of data on cardiovascular disease, which may be utilized to use data mining techniques. In this research five classification techniques have been applied on a secondary dataset to compare the accuracy rate to identify the presence of CVD in human body.

The first part of this paper contains the review of literatures on cardiovascular sickness. In the following part, the reasons for Cardiovascular sickness are researched. Furthermore, the conventional CVD identification approaches and their drawbacks are investigated. From that point forward, the purposes of data mining strategies to predict Cardiovascular illness is proposed. The acquired dataset is then discussed, which was utilized to apply several classification approaches. Moreover, the data analysis is demonstrated after applying classification techniques on the dataset. Then, a comparison of the accuracy of several data mining methods for identifying cardiovascular disease is presented.

2. Cardiovascular Disease

A cardiovascular sickness is a wide-extent of issues that impact the heart and veins. It can also cause abnormalities in the way blood flows throughout the body [3].

Cardiovascular disease (CVD) manifests itself in a number of ways, including illness, disability, and death. In medicine, illness diagnosis is a crucial and difficult task [3]. Universally cardiovascular infection has been the one of the significant reasons for death for a really long time [4]. In 2017, it is estimated that it caused roughly 17.9 million fatalities, accounting for around 15% of all natural deaths [5]. In 2020, it was anticipated that cardiovascular sicknesses will turn into the main source of death and morbidity in emerging countries [6]. By 2030, it is predicted that 23.6 million people would die from CVD if present rates are allowed to continue [7].

The World Health Organization (WHO) evaluated that about 12 million individuals lost their life every year because of heart disease [3]. In the US and other industrialized nations, this illness represents about portion, everything being equal. In different nations, for example, India and Bangladesh, coronary illness is viewed as the main source of death. In every 34 seconds one individual passes on because of coronary illness in the United States [3]. Unhealthy or irregular diet, consuming tobacco, alcohol and drugs are cause of cardiovascular disease. Diet is a vital factor for cardiovascular illness which is controllable [8]. An expansion in lipids, oil, and stoutness is likewise a reason for cardiovascular disorder. Lipids accumulate in the arteries, causing them to compress and decrease blood flow [9]. The arteries get more congested with plaque buildup [10]. The researches informs that the cardiovascular disease is caused by high blood pressure, cholesterol, and a rapid pulse rate [9]. There are also some non-modifiable elements. Age is a major factor that can't be changed and is likewise a reason for cardiovascular sickness. Tobacco use is to blame for 40% of all heart disease fatalities [9]. It damages and tightens blood vessels because it lowers the oxygen content in the circulation [9].

3. Investigation Techniques

In modern medical science, there are different investigation techniques to identify the cardiovascular disease or the heart disease. The most reliably used techniques are Electrocardiogram, Echocardiography, Cardiovascular Magnetic Resonance Imaging, Computed Tomography and Angiogram. These techniques are likewise very fruitful in identifying cardiovascular issues. Usually, these techniques are used whenever men or women face or feel any cardiac abnormality.

3.1. Electrocardiogram and its Flaws

An electrocardiogram (ECG) spreads 12 lead electrodes over human tissue and examines the electrical activity of the human heart to identify anomalies in the cardiac cycle [11]. ECG was developed and standardized using male anatomy and male experience with cardiac disease, therefore using it as the only monitoring modality for women is definitely not acceptable [12]. The 'sex' attribute is the most important factor that influences high accuracy forecasts for heart disease detection [4]. Traditional ECG visual analysis approaches for clinicians are difficult and time-consuming. To spot the errors in an ECG, visual analysis requires highly experience [13]. The ECG packets create unnecessary delay, which is undesirable for individuals with cardiovascular disease (CVD) and the signals of ECG are enormously large in size [14]. When an ECG is used as a main diagnostic tool, it frequently results in poor diagnostic specificity. Furthermore, ECG reporting by general practitioners has a wide range of values, adding to the diagnostic ambiguity [15].

3.2. Echocardiography and its Flaws

There is another technique, which is Echocardiography. according to the ACC/AHA and ESC guidelines, the

single most helpful test in the diagnosis of heart failure is echocardiography. since to lay out a reasonable finding of cardiovascular sickness, underlying irregularity, systolic brokenness, diastolic brokenness, or a blend of these anomalies should be kept in people who present with resting or exertional side effects of cardiovascular breakdown [16]. The 2D strategy and M-mode echocardiogram are utilized to decide the volume and mass of the Left Ventricle. The mass of asymmetric ventricles cannot be correctly approximated since they involve assumptions about the form of the left ventricle [17]. In echocardiography, accurate pericardium measurement is difficult (TEE superior to TTE). Some people (e.g., those with a high BMI or chronic obstructive lung illness) may have narrow echocardiography windows [18].

3.3. Angiogram and its Flaws

Angiogram is the best quality level test for distinguishing the presence and seriousness of atherosclerotic cardiovascular infection. Angiogram also has some complication. Allergic and Adverse Reactions, Nephropathy, Cholesterol Emboli, Vascular Injury even death are the complications of Angiogram [19]. The presence of multivessel disease, left main coronary artery disease (LMCA), CHF, renal insufficiency, and advanced age are the most significant baseline characteristics that affect mortality after coronary angiography [19].

3.4. MRI and its Flaws

Attractive Resonance Imaging (MRI) is a gigantic clever gadget for cardiovascular sickness. Heart attractive reverberation imaging (MRI) is the best quality level for right ventricular evaluation since it is an exceptionally exact and reproducible proportion of right ventricular size and capacity. Not at all like customary 2D echocardiography, cardiovascular MRI considers the recording and translation of a volumetric dataset, which takes into consideration a superior comprehension of right ventricular intricacy [20]. MRI has various drawbacks, including decreased relative availability, expanded securing time, relative price, Certain groupings (for instance, pericardial enhancement) need intravenous gadolinium. Pacemaker patients ought to keep away from this medication [18].

3.5. CMR and its Flaws

There is correspondingly a best quality level test which is Cardiovascular Magnetic Resonance imaging (CMR). Cardiac magnetic resonance (CMR) has developed into a critical technique for the diagnostic and prognostic evaluation of patients with heart failure [21]. CMR imaging is the best quality level for deciding the mass and volume of the left ventricle. Claustrophobia, Artifacts initiated by arrhythmias and patient movement and cut-off points getting from patient size are restrictions of CMR [17]. CMR can be used to recognize the different non-ischemic cardiomyopathy etiologies [21]. In spite of the capability of CMR for the assessment of valve illness, there are still some limitations. Regardless of various distributed approval studies, these procedures have not been widely applied, and there is minimal clinical experience [22]. CMR has long procurement time, not widely available, intravenous gadolinium [18].

3.6. Cardiac Computed Tomography and its Flaws

Heart Computed Tomography (CT) is one more technique to perceive the cardiovascular sickness. The development of electron beam CT (EBCT), which gave appropriate temporal resolution to address the issue of heart motion, revolutionized cardiac computed tomography (CT) imaging. Radiation and iodine contrast openness are two clear disadvantages of cardiovascular CT. The difference in assumptions between the methodologies, as well as the impact of b-blockade before imaging and the influence of contrast, might all contribute to this systematic inaccuracy. The extra coronary angiography data offered by CT is crucial, despite the slight measurement inaccuracy [17]. Cardiac Computed Tomography (CT) also creates associated radiation exposure, which burn the skin [18].

4. Data Mining in Cardiovascular Disease

Alongside customary examination procedures, present day world can distinguish the coronary illness utilizing information mining methods or calculations. Clinical decisions are as a rule made considering expert's association and experience rather than the data set's rich knowledge. This practice results in unintended biases, mistakes, and exorbitant medical expenditures, all of which have an impact on the quality of care offered to patients. Data mining has the ability to create a knowledge rich environment that can enhance the quality of therapeutic judgments dramatically [23]. The act of uncovering previously undiscovered patterns and trends in databases and utilizing that knowledge to develop predictive models is known as data mining [24]. In the field of health care data mining is becoming more common and vital, Table 1.

There is potential for improved and more accurate prediction and detection of cardiac disease by using data mining-based approaches [25]. In shrewd clinical frameworks, data mining has played a fundamental role [26]. In the area of medical, the most widely utilized methods are Naïve Bayesian, Decision tree, Support vector machine, K-Nearest Neighbors, Fuzzy logic, Fuzzy based neural network, genetic algorithms and Artificial neural network [27,28]. The literature reviews make clear that Naïve Bayes, Decision Tree, Support Vector Machine, K-Nearest Neighbors and Random Forest classification techniques are superior classifier to predict human diseases. For this reason, these algorithms were proposed as investigation techniques to predict CVD. However, their performance is data dependent, comparing the accuracy of the strategies and determining the best one is difficult [28]. In Table 1, the most commonly

used data mining classification techniques for predicting CVD are presented, the effectiveness of these techniques and their working strategy are also described.

5. Dataset

In this research, a secondary dataset has been used which was collected from internet (Kaggle.com). In this dataset, there are 303 individuals whose information has been collected. Here, the range of age is from 27 to 77 years old and among them 207 are male and 96 are female. After analysing the dataset, there were 12 attributes which have been selected, Table 2. The selected attributes are age, sex, chest pain (cp), serum cholesterol (chol), fasting blood sugar (fbs), resting blood pressure (trestbps), maximum heart rate (thalach), exercise induced angina (exang), depression induced by exercise relative to rest (oldpeak), number of major vessels colored by flourosopy (caa), the slope of the peak exercise ST segment (slope) and presence of cardiovascular disease (target). From the dataset it has been found that 143 individuals have typical type-1, 50 are in typical type angina, 87 are in non-angina pain and 23 are victim of asymptomatic chest pain.

Table 1. The summary of data mining techniques that are popularly used in medical science.

Classification technique	Summary
Naïve Bayes	It has the ability to tackle diagnostic and predictive issues [29]. It is more preferable algorithm when the traits are independent of one other [30]. It pulls concealed data from a database and compares user values to a trained data set [31]. It can answer difficult questions about heart disease diagnosis, allowing healthcare providers to make more informed clinical judgments than standard decision support systems [31]. It also helps to cut treatment expenses by giving effective remedies [31].
Decision Tree	In order to improve prediction accuracy, this strategy, recursively splits data into branches to form a tree [28]. The success of a decision tree model is determined on the data, although it is generally accurate [32]. The classification report for heart illness is generated using a decision tree technique [9]. Only Decision Trees include the drill through functionality for accessing comprehensive patient profiles [33].
Support Vector Machine	For direct and non-straight information SVM classification approaches is applied [28]. SVM accomplishes classification tasks by increasing the margin between the two categories while reducing classification errors [28]. Some benefits of SVM are it has a strong mathematic base; it has the idea of structural risk reduction that translates into the minimizing of the chance of an erroneous classification on fresh cases [34]. SMO is a Support Vector Machines (SVM) method that has been enhanced by tailored training (SVMs) [35].
K-Nearest Neighbors	K-Nearest Neighbors (KNN) is a straight forward method that stays aware of all events and orders new ones using a similarity measure. Starting around 1970, KNN calculations have been utilized in a few applications, for example, factual assessment and pattern identification [36]. In the dataset initialization KNN is also used. That is why it is called as pre-processed algorithm [9].
Random Forest	In Random Forests, a N-number of choice trees are framed, and every choice tree is picked in view of a vote, after which the class determination is assessed [37]. The random forest algorithm determines which variables are significant in identification and performs well on vast data sets. It can handle a large number of input variables including missing values [38]. It generates very accurate classifications for numerous data sets, particularly the heart disease data set [37].
Neural Networks	Artificial Neural Networks are network structures model after human neurons and they are made up of a number of nodes linked by directed connections, each of which represents a processing unit, and the connections between them show their causal relationship. This categorization approach is becoming a significant tool in data mining, and it may be utilized in descriptive and predictive data mining for a variety of reasons [32]. The Neural Network predicts cardiac illness with the least amount of error [9]. Artificial Neural Networks (ANNs) are utilized in clinical decision-making to assist physicians in analysing and comprehending complicated clinical data and medical applications [32].

Table 2. Selected attributes form the dataset.

Attribute name	Attribute's Value type	Description
age	Numeric	From 29 to 77 (in year)
sex	Nominal	Male (207), female (96)
cp	Nominal	Typical type-1 (143), typical type angina (50), non-angina pain (87), asymptomatic (23)
trestbps	Numeric	94 to 200 (mmHg)
chol	Numeric	126 to 564 (mm/dl)
fbs	Nominal	TRUE ≥ 120 mg/dl FALSE ≤ 120 mg/dl
thalach	Numeric	From 71 to 202
exang	Nominal	Yes (99), no (204)
oldpeak	Numeric	From 0 to 6.2
slope	Nominal	Unsloping (21), flat (140), down sloping (142)
caa	Numeric	From 0 to 4
target	Nominal	True (165), False (138)

According to the dataset, the risk of suffering from cardiovascular disease is significant among 48 to 53 years old persons. From the dataset, we found that women are more likely affected than men are by CVD. People who has typical type-1 chest pain are less affected by CVD but who are suffering from typical type angina, non-angina pain or asymptomatic levels of chest pain at high risk of having CVD, Fig. 2(A). It has been found that People who has no exercise induced angina are at high risk of CVD.

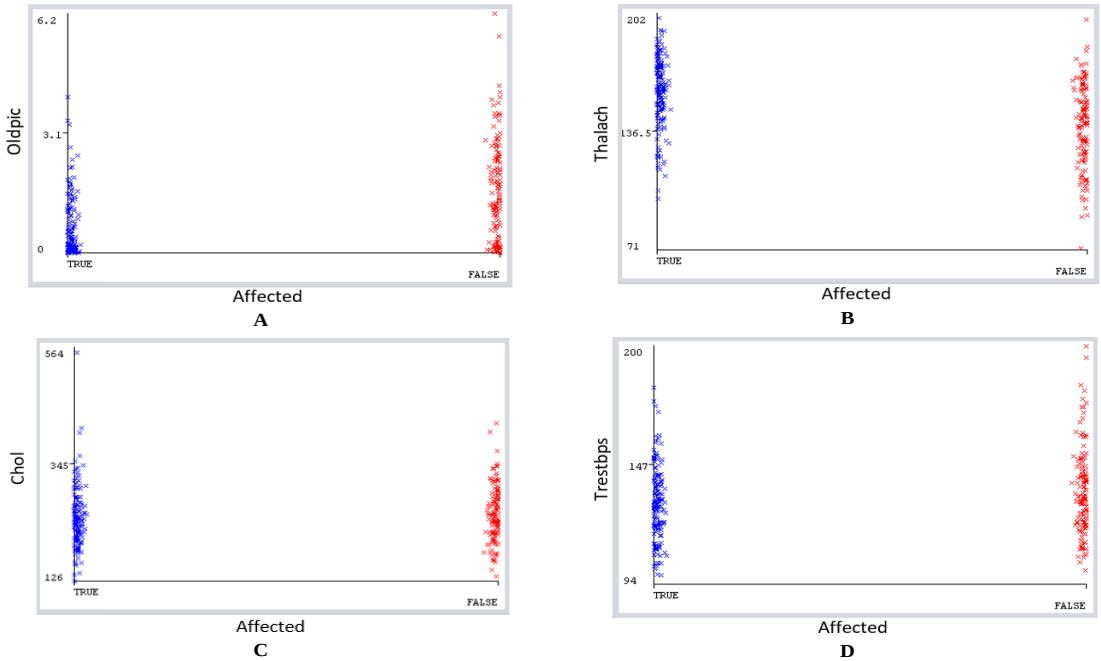

Fig.1. Affected (blue dot) and non-affected (red dot) people by Cardiovascular Disease with the changes of oldpeak (A), thalach (B), chol (C), trestbps (D) attributes value.

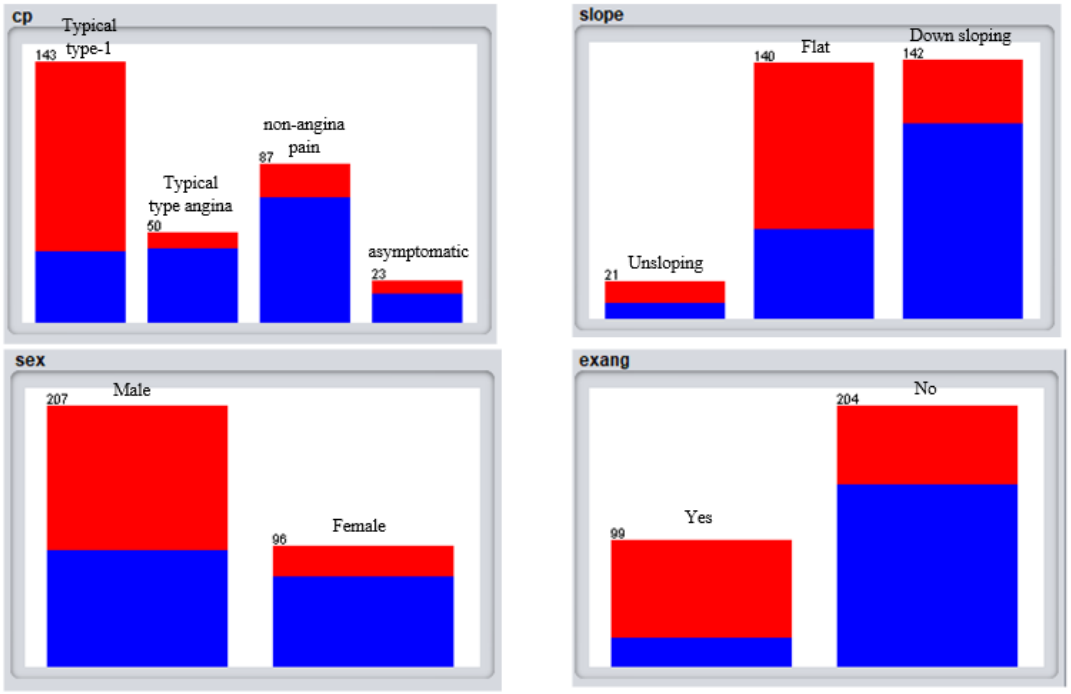

Fig.2. Ratio of affected and non-affected people by based on cp, slope, sex, exang attribute.

6. Result Analysis of the Classification Techniques

After applying classification algorithms using Weka tools, it has been found that the highest accuracy rate of Correctly Classified Instance is found for Naïve Bayes and SMO, which is 83.1683% for both of them, Table 3.

In table 3, Naïve Bayes and SMO both algorithms predicted 252 instances accurately and 51 instances inaccurately

out of 303 instances. However, the accuracy rate remains same but there are some differences in how errors are measured. In Naïve Bayes the Mean Absolute Error and Relative Absolute Error is higher and Root Mean Squared Error and Root Relative Squared Error are comparatively lower than SMO. The other three algorithms that were applied on the dataset also performed really well to predict cardiovascular disease, Table 3. K-Nearest Neighbors successfully predicted 244 instances correctly and 59 instances incorrectly. Random forest correctly predicted 245 instances and incorrectly predicted 58 instances. Lastly, Decision tree correctly predicted 237 instances and incorrectly predicted 68 instances. For predicting CVD, the highest accuracy rate was found by applying Naïve Bayes and SMO which is 83.1683 percent for both of them. The accuracy rate of K-Nearest Neighbors (IBK) and Random Forest are consequently 80.5281 percent and 80.8581 percent. The lowest accuracy rate was found after applying the Decision Tree algorithm which is 78.2178 percent.

Table 3. Accuracy comparison of the used data mining techniques.

	Classifiers Name				
Evaluator/ classifier	Naïve Bayes	K-Nearest Neighbors (IBK)	Random Forest	SMO	Decision Tree (J48)
Total Number of Instance	303	303	303	303	303
Correctly Classified Instance	252	244	245	252	237
Incorrectly Classified Instance	51	59	58	51	68
Kappa Statistic	0.6584	0.6072	0.6109	0.6584	0.5593
Mean Absolute Error	0.2049	0.1968	0.279	0.1683	0.2862
Root Mean Squared Error	0.3747	0.4398	0.3673	0.4103	0.4413
Relative Absolute Error	41.3009%	39.6731%	56.2397%	33.9277%	57.6883%
Root Relative Squared Error	75.2403%	88.2944%	73.755%	82.3726%	88.6045%
Accuracy of Correctly Classified Instance	83.1683%	80.5281%	80.8581%	83.1683%	78.2178%

This represents that with the help of Naïve Bayes and SMO classifier, CVD can be predicted most accurately rather than other three algorithms. As Decision tree's accuracy rate is comparatively lowest, it should be used less when the dataset contains most of the attributes that were present in this research.

6.1. Statistics and Graphs

After applying classification algorithms on the dataset, Weka tool provides some statistics and graph which illustrate a better analytical view.

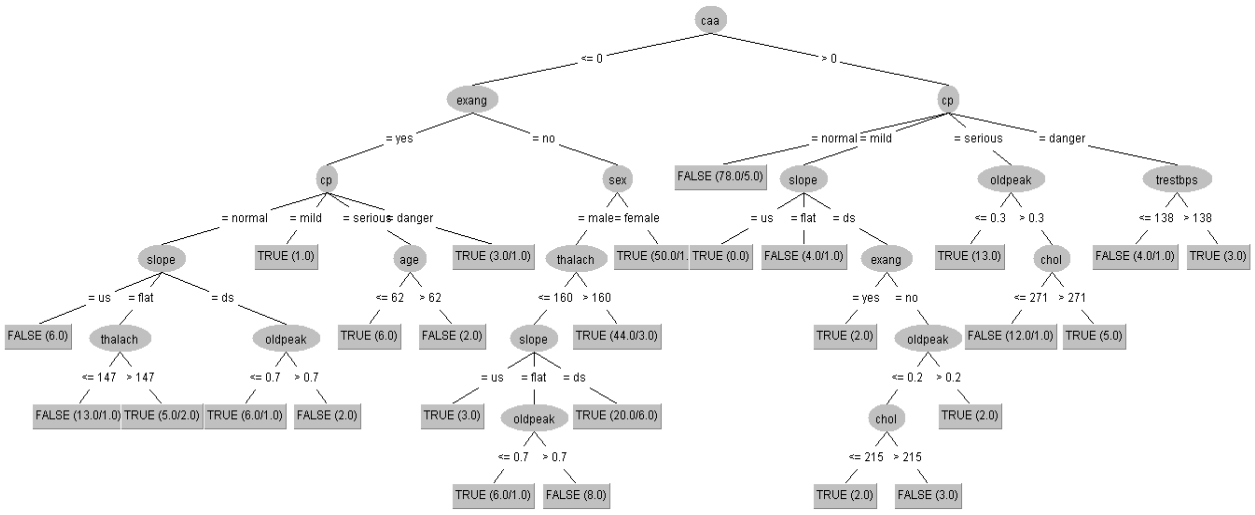

Fig.3. Visualization of decision tree.

It has been found that the risk of CVD is high when the value of oldpeak (depression induced by exercise relative to rest) attribute is less than 2.4 and if the value increases form 2.4 then the risk of CVD is less, Fig. 1(A). It suggests that having a high oldpeak is beneficial for preventing CVD. If the value of thalach (maximum heart rate) is less than 142 then the possibility of the presence of CVD in human body is less, Fig. 1(B). If the thalach's value is greater than 142 then the possibility of CVD is high. This demonstrates that a low thalach value is advantageous for preventing CVD. It has been found that chol attribute is the least important attribute in this dataset to predict cardiovascular disease, Fig. 4. There are no discernible variations in the ratio of CVD affected to non-affected patients as cholesterol levels

fluctuate, Fig. 1(C). So, only depending on the level of chol (serum cholesterol), it is not possible to properly identify the presence of CVD in human body. It has been discovered that caa (number of major vessels colored by flourosopy) is the most effective attribute in this dataset that affects the most to the predict the possibility of CVD, Fig. 4. The possibility of CVD is higher when the caa's value is zero, Fig. 5. The possibility of CVD decreases with caa's value increases and when the value is 4 then the possibility of CVD is very less, Fig 5. As a result, having a high caa is advantageous for avoiding CVD. People with caa is greater than 0, type of cp (chest pain) is in danger (asymptomatic) and trestbps (resting blood pressure) is less than 138 (mmHg) then the possibility of CVD is less, Fig. 1(D). People with trestbps higher than 138 (mmHg) are mostly affected by CVD, Fig. 1(D). In the collected dataset, the ratio of female being affected by CVD is higher than male. For female, even if there is no exang (exercise induced angina) but caa's value is less than or equal 0 then the possibility of being affected by CVD is higher, Fig. 3. It has been discovered that even people with no exercise induced angina are also in risk of CVD. In fact, in this collected dataset the ratio of affected people is higher among them who have no exercise induced angina, Fig. 2. So, people with no exercise induced angina also should be aware of CVD as well as who are already suffering from exang (exercise induced angina). People with normal (typical type-1) level of chest pain are less affected by CVD, Fig. 2. This demonstrate that having typical type-1 angina is comparatively good to be safe from CVD. However, it has been discovered that when the type of cp (chest pain) is in normal (typical type-1) and the value of caa (number of major vessels colored by flourosopy) attribute is greater than 0, then the possibility of CVD is lower, Fig. 3. But the possibility is higher for them who are affected by mild (typical type angina), serious (non-angina pain) and danger (asymptomatic) level of chest pain. The ratio of affected people is also high among these types of chest pain's victims, Fig. 2. This suggests that persons experiencing mild, serious, or dangerous levels of chest pain are most probably suffering from CVD. It has been found that if the slope (peak exercise ST segment) attribute's value is down sloping then the ratio of being affected by cardiovascular disease is higher, Fig. 2. On the other hand, the ratio is comparatively lower when the slope's type is unsloping and flat, Fig. 2. This indicates that having peak exercise ST segment (slope) down sloping increases the risk of CVD.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 12 target):
        Gain Ratio feature evaluator

Ranked attributes:
 0.164612  11 caa
 0.156009   8 exang
 0.132156   7 thalach
 0.117624   3 cp
 0.105318   9 oldpeak
 0.090289  10 slope
 0.065643   2 sex
 0.060273   1 age
 0.000934   6 fbs
 0          4 trestbps
 0          5 chol

Selected attributes: 11,8,7,3,9,10,2,1,6,4,5 : 11
```

Fig.4. Ranking of most effective attribute after applying ranker algorithm.

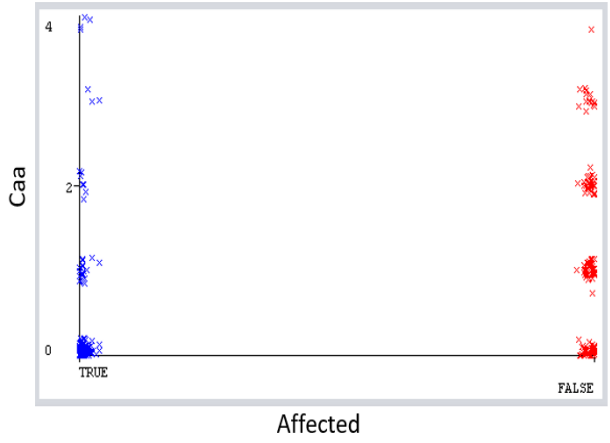

Fig.5. Changes of affected people by CVD based of Caa arrtibute's value.

7. Conclusion

In this research, the main focus was to do comparative analysis of data mining techniques to predict cardiovascular disease. Thus, the classification approaches can give an early prediction of Cardiovascular illness more accurately. Throughout this research, five classification strategies are used which are Naïve Bayes, Support Vector Machine (SMO), Random Forest, Decision Tree and K-Nearest Neighbors (IBK), Table 1. These five techniques are applied on a dataset that contains various attributes to analysis the best accuracy rate, Table 2. After applying these classification techniques, the comparison of accuracy rate has been established. After analysing the data and graphs, the probability and risk factors of CVD were identified depending on various attributes.

In the selected 12 attributes of this dataset, the high value of thalach (maximum heart rate), trestbps (resting blood pressure) and cp (chest pain) are harmful for human body because increases of these attribute's value lead to cardiovascular disease. On the other hand, the high value of oldpeak (depression induced by exercise relative to rest) and caa (number of major vessels colored by flourosopy) are advantageous in avoiding CVD. Among the applied classification techniques, Naive Bayes and SMO provided the best accuracy rate and Decision Tree provided the lowest accuracy rate to predict CVD using this dataset, Table 3.

This result can be use as the reference to give a support to the investigation techniques of CVD. In the near future the researcher can use the Naive Bayes and SMO to get better accuracy rate while working with these 12 attributes and from these two data mining approaches, can pick the better one to apply on large dataset to get more accurate prediction rate of cardiovascular disease. There is a limitation in this research that a secondary dataset has been used because this research was done in covid-19 period. Due to lockdown and restriction, it was not possible to collect primary data. In future, the primary data would be collected from different part of the country and organizations and compare the results with this research finding.

References

[1] Manisha Barman, J Paul Chaudhury, "A Framework for Selection of Membership Function Using Fuzzy Rule Base System for the Diagnosis of Heart Disease," *Information Technology and Computer Science,* vol. 5, num. 11, pp. 62-70, 2013.

[2] Isra'a Ahmed Zriqat, Ahmad Mousa Altamimi, Mohammad Azzeh, "A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods," *International Journal of Computer Science and Information Security (IJCSIS),* vol. 14(12), pp. 868-879, 2016.

[3] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction*," International Journal of Computer Applications (0975 –8887)* vol. 17–No.8, March 2011.

[4] Shafenoor Amin, Mohammad; Kia Chiam, Yin; Dewi Varathan, Kasturi. "Identification of Significant Features and Data Mining Techniques in Predicting Heart Disease." *Telematics and Informatics*, Vol. 36, pp. 82-93, March 2019.

[5] Pratiksha Shetgaonkar, Dr. Shailendra Aswale, "Heart Disease Prediction using Data Mining Techniques," *International Journal of Engineering Research & Technology* (IJERT), ISSN: 2278-0181 IJERTV10IS020083, vol. 10 Issue 02, February-2021.

[6] David S. Celermajer, Clara K. Chow, Eloi Marijon, Nicholas M. Anstey, Kam S. Woo, "Cardiovascular Disease in the Developing World: Prevalences, Patterns, and the Potential of Early Disease Detection," *Journal of the American College of Cardiology*, vol. 60, pp. 1207-1216, 2012.

[7] Vikas Chaurasia, Saurabh Pal, "Data Mining Approach to Detect Heart Diseases," *International Journal of Advanced Computer Science and Information Technology (IJACSIT),* vol. 2, no. 4, pp. 56-66, 2013.

[8] Shilpa N. Bhupathiraju; Katherine L. Tucker, "Coronary heart disease prevention: Nutrients, foods, and dietary patterns," *Clinica Chimica Acta,* vol. 412, pp 1493-1514, 2011.

[9] J. Thomas, Theresa Princy. R, "Human Heart Disease Prediction System Using Data Mining Techniques," *IEEE 2016 International Conference on Circuit, Power and Computing Technologies* (ICCPCT) - Nagercoil, India, 2016.

[10] Obanijesu Opeyemi, Emuoyibofarhe O. Justice, "Development of Neuro-fuzzy System for Early Prediction of Heart Attack," *Information Technology and Computer Science,* vol. 4, num. 9, pp. 22-28, 2012.

[11] M. Vijayavanan, V. Rathikarani, Dr. P. Dhanalakshmi, "Automatic Classification of ECG Signal for Heart Disease Diagnosis using Morphological Features," *International Journal of Computer Science & Engineering Technology* (IJCSET)," ISSN: 2229-3345, vol. 5, No. 04, pp. 449-455, 2014.

[12] Borejda Xhyheri; Raffaele Bugiardini, "Diagnosis and Treatment of Heart Disease: Are Women Different from Men?" *Progress in Cardiovascular Diseases,* vol. 53, pp. 227–236.2010.

[13] Serkan Kiranyaz, Turker Ince, Jenni Pulkkinen, Moncef Gabbouj, "Personalized Long-term ECG Classification: A Systematic Approach," *Expert Systems with Applications,* vol. 38, pp. 3220–3226, 2011.

[14] Fahim Sufi, Ibrahim Khalil, "Diagnosis of Cardiovascular Abnormalities from Compressed ECG: A Data Mining-Based Approach," *IEEE Transactions on Information Technology in Biomedicine,* vol. 15, pp. 33-39, 2011.

[15] Dimitri Grün, Felix Rudolph, Nils Gumpfer, Jennifer Hannig, Laura K. Elsner, Beatrice von Jeinsen, Christian W. Hamm, Andreas Rieth3, Michael Guckert, Till Keller, "Identifying Heart Failure in ECG Data with Artificial Intelligence—A Meta-Analysis," *Frontiers in Digital Health*, February 2021 | Volume 2 | Article 584555.

[16] J. K. Oh, "Echocardiography in heart failure: Beyond diagnosis," *European Journal of Echocardiography,* vol. 8, pp. 4-14, 2007.

[17] Vasiliki V. Georgiopoulou MD, Andreas P. Kalogeropoulos MD, Paolo Raggi MD, Javed Butler MD, MPH, "Prevention, Diagnosis, and Treatment of Hypertensive Heart Disease," *Cardiology Clinics*, vol. 28, pp. 675– 691, 2010.

[18] John D. Groarke1, Paul L. Nguyen, Anju Nohria, Roberto Ferrari, Susan Cheng and Javid Moslehi, "Cardiovascular

complications of radiation therapy for thoracic malignancies: The Role for Non-invasive Imaging for Detection of Cardiovascular Disease," *European Heart Journal*, vol. 35, pp. 612–623, 2014.
[19] Morteza Tavakol MD, Salman Ashraf MD & Sorin J. Brener MD, "Risks and Complications of Coronary Angiography: A Comprehensive Review," *Global Journal of Health Science*, vol. 4, No. 1, pp. 65-93, 2012.
[20] Sushil A. Luis, Patricia A. Pellikka, "Best Practice & Research Clinical Endocrinology & Metabolism," Division of Cardiovascular Diseases, Mayo Clinic College of Medicine, Rochester, MN 55905, USA.
[21] Jorge A. Gonzalez, Christopher M. Kramer, "Role of Imaging Techniques for Diagnosis, Prognosis and Management of Heart Failure Patients: Cardiac Magnetic Resonance," *Springer Science+Business Media (Springer)*, vol. 12(4), pp. 276-283, 2015 doi: 10.1007/s11897-015-0261-9
[22] MD Peter J. Cawley, MD Jeffrey H. Maki, MD Catherine M. Otto, "Cardiovascular Magnetic Resonance Imaging for Valvular Heart Disease: Technique and Validation," *Circulation*, vol. 119(3), pp. 468-478, doi: 10.1161/CIRCULATIONAHA.107.742486
[23] Aqueel Ahmed, Shaikh Abdul Hannan, "Data Mining Techniques to Find Out Heart Diseases: An Overview," *International Journal of Innovative Technology and Exploring Engineering* (IJITEE), ISSN: 2278-3075, Volume-1, 2012.
[24] Bhatla, Nidhi; Jyoti, Kiran, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques," *International Journal of Engineering Research & Technology* (IJERT), vol. 1, ISSN: 2278-0181 ,2012.
[25] Mudasir M Kirmani, "Cardiovascular Disease Prediction Using Data Mining Techniques: A Review," *Oriental Journal of Computer Science & Technology*, vol. 10, pp. 520-528, 2017.
[26] K. Srinivas, G. Raghavendra Rao, A. Govardhan, "Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions using Data Mining Techniques," *IEEE Transaction on Computer Science and Education* (ICCSE), pp. 1344 - 1349, 2010.
[27] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, "Effective Heart Disease Prediction using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542-81554, 2019.
[28] Milan Kumari, Sunila Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction 1," *IJCST*, vol. 2, ISSN: 2229- 4333(Print) | ISSN: 0976-8491(Online), 2011.
[29] Dhanashree S. Medhekar, Mayur P. Bote, Shruti D. Deshmukh, "Heart Disease Prediction System using Naive Bayes," *International Journal of Enhanced Research in Science Technology & Engineering*, vol. 2, ISSN NO: 2319-7463, 2013.
[30] Mrs. G. Subbalakshmi, Mr. K. Ramesh, Mr. M. Chinna Rao, "Decision Support in Heart Disease Prediction System using Naive Bayes," *Indian Journal of Computer Science and Engineering* (IJCSE), vol. 2, pp. 170-176, 2011.
[31] Shadab Adam Pattekari and Asma Parveen, "Prediction System for Heart Disease Using Naive Bayes," *International Journal of Advanced Computer and Mathematical Sciences*, vol. 3, pp. 290-294, 2012.
[32] Umair Shafique, Fiaz Majeed, Haseeb Qaiser, and Irfan Ul Mustafa, "Data Mining in Healthcare for Heart Diseases," *International Journal of Innovation and Applied Studies*, vol. 10, pp. 1312-1322, 2015.
[33] Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar, "Early Heart Disease Prediction Using Data Mining Techniques," *CCSEIT*, DMDB, ICBB, MoWiN, AIAP – 2014, pp. 53–59, 2014.
[34] Fabio Mendoza Palechor*, Alexis De la Hoz Manotas, Paola Ariza Colpas, Jorge Sepulveda Ojeda, Roberto Morales Ortega, Marlon Piñeres Melo, "Cardiovascular Disease Analysis Using Supervised and Unsupervised Data Mining Techniques," *Journal of Software*, vol. 12, 2017.
[35] Wan Hajarul Asikin Wan Zunaidi, RD Rohmat Saedudin, Zuraini Ali Shah, Shahreen Kasim,Choon Sen Seah and Maman Abdurohman, "Performances Analysis of Heart Disease Dataset using Different Data Mining Classifications," *International Journal on Advanced Science Engineering and Information Technology*, vol. 8, pp. 2677-2682, 2018.
[36] M.Akhil jabbar, B.L Deekshatulu, Priti Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," *Procedia Technology*, vol. 10, pp. 85-94, 2013.
[37] Indu Yekkala, Sunanda Dixit, "Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection," *International Journal of Big Data and Analytics in Healthcare*, vol. 3, pp. 1-12, 2018.
[38] M. A. Jabbar, B. L. Deekshatulu and Priti Chandra, "Intelligent Heart Disease Prediction System using Random Forest and Evolutionary Approach," *Journal of Network and Innovative Computing*, vol. 4, pp. 175-184, 2016.

Authors' Profiles

Md. Al Muzahid Nayim has received his B.Sc. in Computer Science and Engineering at the Faculty of Science and Technology from American International University-Bangladesh (AIUB), 2022. His major was Software Engineering. Currently working as Junior Software Engineer at Incevio, Dhaka, Bangladesh. His research interest includes Data Mining, Data Warehouse, Deep Learning, and AI.

Fahmidul Alam has received his B.Sc. in Computer Science and Engineering at faculty of Science and Technology from American International University-Bangladesh (AIUB), 2022. His major was Information System. His research interest includes Data Mining, Robotics and Image Processing.

Md. Rasel has received his B.Sc. in Computer Science and Engineering at the Faculty of Science and Technology from American International University-Bangladesh (AIUB), 2022. His major was Software Engineering. His research interest includes Data Mining, Data Warehouse, and Machine Learning, Block Chain.

Ragib Shahriar received his B.Sc. in Computer Science and Engineering at the Faculty of Science and Technology from American International University-Bangladesh (AIUB) in 2022. His major was Information Technology. He is currently working as Junior Software Engineer at INovex Idea Solution, Dhaka, Bangladesh. His research interest includes Data Mining, Data Warehouse, and Data Science.

Dr. Dip Nandi currently works as a Professor and the Director of Faculty of Science and Technology in American International University-Bangladesh (AIUB). DR. Nandi achieved his Doctor of Philosophy (PhD) degree from RMIT, Australia and MSc degree from The University of Melbourne, Australia. His research area includes: Software Engineering, E-Learning Technologies, Data Mining and Information systems. He has supervised more than 70 students as thesis supervisor. DR. Nandi is associated with several organizations such as IEEE, ACM. He has published several peer-reviewed journal articles.

How to cite this paper: Md. Al Muzahid Nayim, Fahmidul Alam, Md. Rasel, Ragib Shahriar, Dip Nandi, "Comparative Analysis of Data Mining Techniques to Predict Cardiovascular Disease", International Journal of Information Technology and Computer Science(IJITCS), Vol.14, No.6, pp.23-32, 2022. DOI:10.5815/ijitcs.2022.06.03