Diagnosis of Coronary Artery Diseases using Classification Algorithms based on Wavelet Transforms

Rajesh Kumar T., Srinivasa rao A., Ashok B., Rajesh Kumar E.

Abstract: ***One of the primary drivers of the death in the world is Coronary Artery Diseases (CAD) which is a major threat in developing and developed countries. The fundamental drivers in CAD leads to blockage of the coronary lumen subsequently blood clot and that prompts to damage of heart muscles or unexpected heart attack which causes death. It is difficult to ascertain that a certain person has been affected by CAD, since there are bunch of parameters has been involved to ascertain the conclusion. Classification has been done using wavelet transform to classify the certain parameters. We analyzed following methods such as NB, Logistic, SMO, RBF Network, K-star, Multiclass Classifier, Conjunctive rule, Decision table, LMT, NB Tree, DTNB, LAD Tree, Random Tree and Random Forest calculations has been associated with extensive fragment of the surveys. This database has been generated from UCI machine learning database. In this paper, we used k-fold cross validation with k values as 10, with 14 properties and calculations of Accuracy, Precision, TPR, FPR, Recall, F-measure and ROC are analyzed practically. The experimental evaluation shows the improvement in accuracy rate of 77.0%, by using the Logistic, SMO and LMT algorithms than the traditional method.***

Index Terms**: *Heart disease, Wavelet transform, Haar WT, Coronary Artery Diseases(CAD)***

I. INTRODUCTION

Natural disasters and accidents causes less amount of death compared with the ethical quality rate of the death caused due to the diseases. According to WHO, it is estimated that, the cardiovascular disease causes 17 million deaths in worldwide[1]. In which coronary artery diseases(CAD) is the major one causes death of 7 million deaths throughout the world per year [1]. Knowledge extraction from a dataset is said to be the data mining. Otherwise, it is a process of using intelligent method of extraction of knowledge from a set of data [2]. The method of decision making in identifying CAD, is called as Angiography.

Revised Manuscript Received on July 06, 2019.

Rajesh Kumar T., Department of CSE, K L E F, Vaddeswaram, Guntur, Andhra Pradesh, India.
Srinivasa Rao A., Department of CSE, K L E F, Vaddeswaram, Guntur, Andhra Pradesh, India.
Ashok B., Department of CSE, K L E F, Vaddeswaram, Guntur, Andhra Pradesh, India.
Rajesh Kumar E., Department of CSE, K L E F, Vaddeswaram, Guntur, Andhra Pradesh, India.

Since the Angiography is costly and high risky factor for the patients, the data mining come into existence and get the researchers attention though the Angiography identifies the position and coverage of the stenotic arteries. Furthermore, Cost-sensitive algorithms can have enormous value in this ground as misclassification of diseased or healthy patients has different costs. An 80% of precision rate of CAD is achieved in Pedreira *et al.*, [3] paper, where they used a neural system on UCI datasets[4]. An accuracy rate of 89.01% was achieved in Das *et al.*[5], where they applied Neural Network as a dataset of Cleveland. A precision rate of 79.17% was achieved by Babaoglu*et al.* [6] using the Support Vector Machine (SVM) Algorithm as dataset. To identify coronary artery diseases, a Fuzzy Model was used by Tsipouras*et al.* [7]. To conclude the information about CAD, Itchhaporia *et al.* [8], used the Neural Network. The reason for the present review is to utilize information mining procedures, which is a managed learning calculation, in order to recognize CAD patients from solid people. The motivation behind the present review is to utilize arrangement calculations to be specific, NB, Logistic, SMO, RBF Network, Kstar, Multiclass Classifier, Conjunctive lead, Decision table, DTNB, LMT Tree, LAD, Random woodland, NB Tree and Random Tree methods. The informational index is obtained from information mining storehouse of college of California, Irvine (UCI)[4]. At last the framework is approved utilizing informational indexes from Long Beach, Switzerland, Hungarian, and Cleveland from Ipoh Specialist Hospital, Malaysia.

A. Wavelet Transform:

A linear signal processing that applied to a data vector X, which then transform to numerical different vector X',without disturbing the energy of the signal is called as the Discrete Wavelet Transform (DWT). Generally, the size reduction idea, that is a multidimensional statistical approach for storing compressed approximated data with minimum missing of information is also discrete wavelet transform. A compressed approximated data means user-identified threshold wavelet value can be stored at a rate of little fraction and making other data as zero. Also, the technique of wavelet transform performs very well to eliminate the noise and irregular data

without affecting the main features of data[11]. If the input vector quantity is having a length of 'n' means, then the time complexity of the algorithm of DWT is *O(n)* in worst case. If we are handling maximum dimensional data, then the DWT is more suitable to use. The fig.1 shows how wavelet transformation is applied for a sample set of four data.

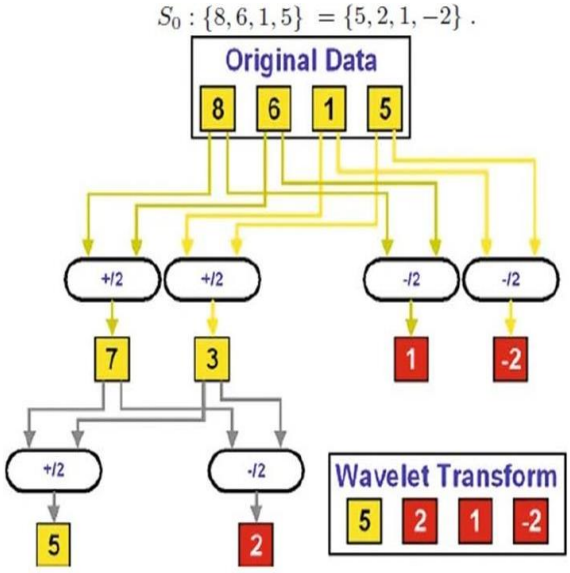

Fig.1. Wavelet transformation of four data points.

B. Haar Wavelet Transform

A Haar wavelet transform is the simplest type of wavelet[11,12]. A mathematical procedure in Haar wavelet transform is called as Haar Transform in discrete signal processing. For all other wavelet transforms, the Haar transform acts as a model. An array in haar transform is divided into two halves from its original length. A running average is first half and running difference is second half. The performance of haar transform is, in a pair of values one half is average and other is difference.

C. *Calculation* Procedure using Haar Transform:

1. For each pair of samples, find the average. (n/2 averages)
2. From the average samples, find the differences between each average. (n/2 differences)
3. Fill the first half of the array with averages.
4. Fill the second half of the array with differences.
5. Repeat the procedure on the first half of the array.

The rest of this paper is organized as follows: Section 2 describes about medical database for heart diseases problem. The section 3 presents about the methods used in data mining. In 4th section, experimental results of all the classification algorithms are discussed and finally section 5 conclude this paper.

II. HEART DISEASE DATABASE

In this paper, the fourteen algorithms namely NB, Logistic, SMO, RBF Network, Kstar, Multiclass Classifier, Conjunctive rule, Decision table, DTNB, LMT, AD Tree, NB Tree, Random tree and Random forest as tested in a medical datasets for Heart disease database obtained from online University of California, Irvine repository (https://archive.ics.uci.edu/ml/datasets/Heart+Disease) before applying it on wavelet transforms. This data set contains A 303 patients with many efficient 54 features are introduced by this dataset. To classify Coronary Artery Disease patients, a gray scale features is employed for left ventricle echo cardio graphics by Acharya et al [13].

A. Data Set Description.

The heart disease of Coronary Artery Disease datasets are availed from University of California, Irvine (UCI) - Data Mining Repository[4]. From Hungarian, Cleveland, Switzerland and VA long beach, the dataset of coronary artery disease of 920 instances were collected, which identifies the CAD. Among these instances, 14 attributes are selected for CAD data. The list of 14 attributes are given below with descriptions which are used in the datasets. 1. age: Age in years, 2.sex: sex (1 = male; 0 = female, 3.cp: Value 1: typical angina, Value 2: atypical angina, Value 3:non-anginal pain,Value 4: asymptomatic, 4.chol: serum cholestoral in mg/dl, 5.fbs:(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false), 6.restecg: resting electrocardiographic results, 7.thalach: maximum heart rate achieved, 8. exang: exercise induced angina (1 = yes; 0 = no, 9. oldpeak = ST depression induced by exercise relative to rest, 10. slope: the slope of the peak exercise ST segment, 11. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect, 12. Ca: Count of fluoroscopy colours in major vessels, 13. num: number of heart disease identification (angiographic disease status, 14. trestbps: resting blood pressure (in mm Hg on admission to the hospital).

Cleveland Data: Dr.Robert Detrano, collected Cleveland data set from VA Medical centre.[4]

Hungarian Data: Dr.AndrasJanosi, collected data set from Hungarian Institute of Cardiology, Budapest. The Cleveland dataset is similar to Hungarian dataset format. The availability of heart disease is 37.5% and Non-availability is 62.5% are there in Class distributions[4].

Switzerland Data: Dr. William Steinbrunn, collected dataset from University Hospital, Zurich, Switzerland. The major count of missing data are related to four datasets of CAD is from Switzerland datasets. The availability of heart disease is 93.5% and Non-availability is 6.5% are there in Class distributions out of 123 instances[4].

III. METHODOLOGY

In this paper, we have tested eleven tree based algorithms namely, NB, Logistic, SMO[9], RBF Network, K-star, Multiclass Classifier, Conjunctive rule [9,10], Decision table, DTNB, LMT, LAD Tree, NB Tree [9], Random tree [9] and Random forest [9] techniques. We calculated Precision, False positive, True positive, F-measure Recall and ROC and selected above eleven classification algorithms to find the highest accuracy rate for cardiovascular diseases (CAD) database. In this experimental methodology, for the input datasets, It support all the mining process to get valid and clear visualization with accuracy results, for 14 attributes with 10 fold cross validation was applied.

IV. PERFORMANCE ANALYSIS AND DISCUSSIONS

In this paper, we used experimental performance result analysis namely, NB, Logistic, SMO, RBF Network, Kstar, Multiclass Classifier, Conjunctive rule, Decision table, LMT, LAD Tree, DTNB , NB Tree, Random tree and Random forest techniques are compared based on the application in the heart diseases database. Weak data mining tool is being used for research banking, education, weather and real database. In this experimental analysis, for the given dataset, we have tested 10 fold cross validation with 14 attributes.

A. Performance Measure

In the field of medicine, especially in heart diseases the accuracy, true positive, true negative, precision and F-measure are to be found with high importance. Subsequently, the accuracy, true positive, true negative, precision and F-measure are used for measuring the performance of algorithms.

B. Confusion matrix

A classification system, having the information of known class and predicted class is generally depicted in a confusion matrix. The results of confusion matrix is displayed in table.1.

Table-1 Confusion matrix

Known Class	Predicted class	
	A	B
A	True positive (TP)	False Negative (FN)
B	False positive (FP)	True Negative (TN)

1. For positive instance, TP is the count of exact predictions.
2. For positive instance, FP is the count of wrong predications.
3. For negative instance, FN is the count of wrong predications.
4. For negative instance, TN is the count of exact predications.

C. True positive and True negative

The ratio of exactly diagnosed Coronary Artery Diseases and sum of the normal samples[9] are called as True positive and Negative.

$$True\ Positive = \frac{TP}{(TP + FN)}$$

$$True\ Negative = \frac{TN}{(TN + FP)}$$

Accuracy

Accuracy is directly proportional to the sum of count of exact predictions to the sum of count of exact and wrong predictions. Next coming equation shows the accuracy[9].

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

Precision

The ratio of the count of significant records to the sum of the count of significant and insignificant records retrieved is said to be precision in percentage[9].

$$Precision = \frac{TP}{(TP+FP)}$$

Recall

The ratio of the count of significant records retrieved to the sum of the count of significant record in dataset is said to be Recall in percentage[9].

$$Recall = \frac{TP}{(TP+FN)}$$

F-measure

Mathematically F-measure is calculated by the harmonic average of both precision and recall.

$$F - measure = \frac{(Precision * Recall)}{(Precision + Recall)}$$

Receiver operating Characteristics (ROC)

A plot drawn between True Positive on y-axis and False Positive on x-axis is said to be the Receiver Operating Characteristics.

Table-2 Confusion matrix and 10-fold cross validation with 14 attributes

Confusion Matrix			Methods	TP Rate	FP Rate	Precision	Recall	F-measure	ROC	Accuracy
	AP	AN	Naive Bayes	0.76	0.30	0.75	0.76	0.76	0.81	0.76
AP	422	78								
AN	104	164								
AP	440	60	Logistic	0.77	0.32	0.76	0.77	0.76	0.83	0.77
AN	115	153								
AP	449	51	SMO	0.77	0.33	0.76	0.77	0.76	0.72	0.77
AN	123	145								
AP	434	66	RBF Network	0.75	0.34	0.74	0.75	0.74	0.78	0.75
AN	123	145								
AP	407	93	KStar	0.69	0.41	0.68	0.69	0.69	0.71	0.69
AN	144	124								
AP	440	60	MultiClass Classifier	0.77	0.32	0.76	0.77	0.76	0.83	0.77
AN	115	153								
AP	385	115	Conjunctive Rule	0.68	0.38	0.68	0.68	0.68	0.69	0.68
AN	125	143								
AP	405	95	Decision Table	0.71	0.37	0.70	0.71	0.70	0.77	0.71
AN	126	142								
AP	415	85	DTNB	0.73	0.34	0.73	0.73	0.73	0.79	0.73
AN	116	152								
AP	415	85	LAD Tree	0.74	0.33	0.73	0.74	0.73	0.78	0.74
AN	114	154								
AP	445	55	LMT	0.77	0.32	0.77	0.77	0.76	0.83	0.77
AN	114	154								
AP	409	91	NB Tree	0.73	0.33	0.73	0.73	0.73	0.78	0.73
AN	112	156								
AP	417	83	Random Forest	0.74	0.32	0.74	0.74	0.74	0.81	0.74
AN	110	158								
AP	373	17	Random Tree	0.68	0.37	0.68	0.68	0.68	0.68	0.68
AN	118	150								

The Table-2 shows the Confusion matrix and 10 fold cross validation with 14 attributes. The plot is also drawn for the above table's data. Thus the fig.2 stipulates the comparison of 14 classification algorithm. Finally, after comparison of all the algorithm, the three algorithms such as Logistic, SMO and LMT algorithms provide the highest accuracy rate of 77.0% which has been experimentally proved.

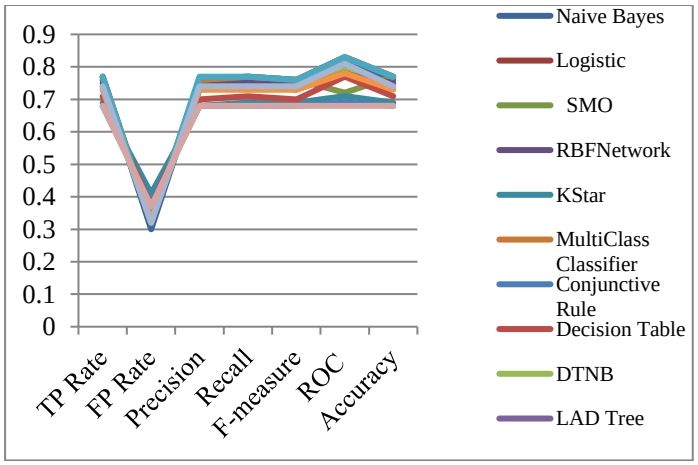

Fig.2. Comparison of 10-fold cross validation with 14 attributes

V. CONCLUSION

In this paper, classification of 14 attributes display in view of discrete wavelet transform was exhibited. As a dimension minimizing function, we utilized Haar Wavelet Transform. A maximum accuracy rate of 77% is obtained using UCI datasets for identifying Coronary Artery Diseases by employing the Logistic, SMO and LMT algorithms in 10 fold cross validation with 14 attributes. In future research, a large database with more attributes will be used to achieve better results.

REFERENCES

1. Bonow R.O, Mann D., Zipes D. P., and P. Libby P, "Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine", 9th edition: New York, Saunders, 2012.
2. Bickel S and Scheffer T, "Multi-view clustering". In Proc. of the IEEE International Conference on Data Mining, pp. 19–26, 2004.
3. Pedreira C.E, Macrini L, and Costa E.S, "Input and Data Selection Applied to Heart Disease Diagnosis", Proc. of International Joint Conference on Neural Networks, IEEE, 2005.
4. (https://archive.ics.uci.edu/ml/datasets/Heart+Disease)
5. Das R., Turkoglu I, Sengur A., "Effective diagnosis of heart disease through neural networks ensembles", Expert Systems with

Applications, pp. 7675–7680, 2009.
6. Babaoglu I, Fındık O., and Bayrak M., "Effects of principle component analysis on assessment of coronary artery diseases using support vector machine", Expert Systems with Applications, pp. 2182–2185, 2010.
7. Tsipouras M., Exarchos T., Fotiadis D., Kotsia, Vakalis A.K.,Naka K, Michalis L., "Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modelling", IEEE Transactions on information technology in biomedicine , Vol.12, NO.4, pp.447-458, 2008.
8. Itchhaporia D., Almassy R., Kaufman L., Snow P., and Oetgen W., "Artificial neural networks can predict significant coronary disease", J.Am. Coll. Cardiol., Vol.28, NO.2, pp.515-521, 1995.
9. Ganesan P., Siva S.K, and Sundar.S, An Experimental Analysis of Classification Mining Algorithm For Coronary Artery Disease, "International Journal of Applied Engineering Research", Volume 10, Number 6 (2015) pp. 14467-14477.
10. Ganesan P, Siva S.K., and Sundar S., A Comparative Study on MMDBM Classifier Incorporating Various Sorting Procedure ," Indian Journal of Science and Technology " Vol 8(9), 868–874, May 2015.
11. James S. Walker. 1999. A Primer on Wavelets and Scientific Applications. Jiawei Han, MichelineKamber Data mining: concepts and techniques: Second Edition illustrated. Morgan Kaufmann Publishers, Inc, 2006.
12. Kiran K.R, Ali MirzaMahmood, MrithyumjayaRao K., "Generating Optimized Decision Tree Based on Discrete Wavelet Transform", International Journal of Engineering Science and Technology Vol. 2(3), 2010, 157-164.
13. Ben-Hur A., and Weston J., "A User's Guide to Support Vector Machines", Methods in Molecular Biology, 2010, pp.223-239.

AUTHORS PROFILE

Rajesh Kumar T. received his bachelor of Engineering degree in Electronics and Communication Engineering from Madras University during 1996, Master of Engineering degree in Computer Science and Engineering from Manonmaniam Sundaranar University during 2004 and thesis submitted for Ph.D program in faculty of information and Communication Engineering under the guidance of Dr.Suresh at Anna University, Chennai during 2019. He is currently working as Assistant Professor in Computer Science and Engineering at Koneru Lakshmaiah Education Foundation(Deemed to be University), Guntur, Andhra Prasesh He has more than 18 years of teaching experience in under graduate and master level. He has published more than 15 research papers in journals and conferences. His research area includes Speech/Image Signal Processing, Data mining, Embedded System and Knowledge Engineering. He is a life member of ISTE, IAEng and ACM.

Srinivasa Rao A. received his bachelor of Engineering degree in Computer Science and Engineering from Acharya Nagarjuna University during 2011, Master of Engineering degree in Computer Science and Engineering from Acharya Nagarjuna University during 2014 and seeking for supervisor to do Ph.D program in the faculty of Computer Science and Engineering at Andhra University. He is currently working as Assistant Professor in Computer Science and Engineering at Koneru Lakshmaiah Education Foundation(Deemed to be University), Guntur. He has more than 4 years of teaching experience in under graduate and master level. He has published more than 2 research papers in journals and conferences. His research area includes Data mining, Internet of Things and Knowledge Engineering. He is a life member of IAEng .

Ashok B. received his bachelor of Engineering degree in Computer Science and Engineering from JNTU Kakinada University during 2010, Master of Engineering degree in Computer Science and Engineering from JNTU Kakinada University during 2013 and pursuing Ph.D program in the faculty of Computer Science and Engineering under the guidance of Dr.Sumathi Ganesan at Annamalai University, Chidambaram,Tamilnadu. He is currently working as Assistant Professor in Computer Science and Engineering at Koneru Lakshmaiah Education Foundation(Deemed to be University), Guntur. He has more than 5 years of teaching experience in under graduate and master level. He has published more than 4 research papers in journals and conferences. His research area includes Medical Image Processing, Internet of Things, Data Mining and Data Warehousing. He is a life member of CSI, IAEng.