A Proposed Decision Tree Classifier for Atherosclerosis Prediction and Classification

Yousef K Qawqzeh[1], Mohammad Mahmood Otoom[1], Fayez Al-Fayez[1], Ibrahim Almarashdeh[2], Mutasem Alsmadi[2], Ghaith Jaradat[3]

[1]Computer science department, College of Science, Majmaah University, Saudi Arabia
[2]MIS Dept., College of Applied Studies and Community Service, Imam Abdurrahman Bin Faisal University, Al-Dammam, Saudi Arabia
[3]Department of Computer Science, Faculty of Computer Science and Information Technology, Jerash University, Jordan

Abstract
Cardiovascular diseases (CVD) represent a big threat to human lives. As most of CVD symptoms are developed silently inside our cardiovascular system, the prediction of the disease before it comes to threaten human's life, represents an appreciated achievement. The tracking of the silent development of atherosclerosis inside arteries may yield to new methods for early detection and prevention of CVD. Atherosclerosis is one of the main causes of CVD. The more silent atherosclerosis is, the more difficult to be detected. It represents a chronic disease that causes arterial wall to be stiffen. Normally, people are not visiting a diagnostic center nor consulting their doctor, unless the risk reaches high level. This study utilized features extracted from photoplethysmogram (PPG) for tracking and evaluating the high-risk atherosclerosis. A sample of 196 participants are enrolled in this study. Their carotid intima-media thickness (CIMT) test were recorded. The PPG's indices along with Age index are fed to a decision tree classifier developed in MATLAB to predict and classify new data into high-risk atherosclerosis or normal atherosclerosis. The developed classifier showed promising results in which it revealed an overall accuracy of 82.6%. Additionally, it showed a sensitivity of 89.3% and specificity of 69.2%. These results represent a new possible method to be valid surrogate measure for atherosclerosis along with the used CIMT test.
Key words:
Atherosclerosis; Photoplethysmogram; Classification; Decision tree; Prediction.

1. Introduction

Cardiovascular diseases (CVD) represent a big threaten to human lives. As most of CVD symptoms are developed silently inside our cardiovascular system, the prediction of the disease before threatens human's life, will be an appreciated advancement. The tracking of the silent development of atherosclerosis inside arteries may yield to new methods for early detection and prevention of CVD. Atherosclerosis is one of the main causes of CVD. The more silent atherosclerosis is, the more difficult to be detected. It represents a chronic disease that causes arterial wall to be stiffen. Normally, people are not visiting a diagnostic center nor consulting their doctor, unless the risk reaches high level. This study utilized features extracted from photoplethysmogram (PPG) for tracking and evaluating the high-risk atherosclerosis. PPG is an optical volumetric measure of an organ. It can be defined as blood volume changes inside arteries (Qawqzeh et al., 2015). A sample of 196 participants were enrolled in this study in which their carotid intima-media thickness (CIMT) test and their PPG data were recorded. Table 1 below illustrates the descriptive analysis of participants. The strategies of classification are widely utilized in clinical settings for predicting patient's health status. This work implements a decision tree method to predict and track atherosclerosis accumulation. Several comparative studies showed that decision tree classifiers are simple and accurate (Latha & Jeeva, 2019). The proposed prediction model is constructed using decision tree method in MATLAB environment.

1.1 Data collection methods

This section provides a brief detail about PPG data and CIMT test recordings. A customized PPG setup is used to record PPG data from each participant in a temperature-controlled room ±25o inside an equipped hospital room (General Hospital of Zulfi, Riyadh, KSA). Subjects are asked to be quite for 3 minutes to allow cardiovascular stabilization. A PPG probe is then applied to the right-hand index finger. The patient is asked to remain quiet and breathe normally. PPG recording, for each subject, ran for 2 minutes. Pre-processing such as down-sampling, and de-trending to remove outliers and drifts have taken place. PPG signals are filtered using the band-pass filter (0.6–15 Hz) any respiratory rhythm and higher frequency disturbances. Finally, the extracted features are saved in an Excel sheet for further possible analysis. The following four indices, b/a, RI, SPt, & DiP, were very significant with CIMT test. In addition, 'Age' index was also very significant.

The CIMT data is collected using carotid duplex ultrasound scanner that contains a screen for video display, computer console, and transducer (probe). The recordings were

Manuscript received December 5, 2019
Manuscript revised December 20, 2019

achieved by a specialized Cardiology doctor in general hospital of Zulfi. Table 2 below examines the correlation between the independent variables and the dependent variable. While Table 3 explains the attributes in this model.

Table 1: Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
b/a	196	.49	.87	.6687	.08170	.007
Age	196	18.00	75.00	40.4643	16.18812	262.055
RI	196	.49	.87	.6902	.08581	.007
CIMT	196	.40	.90	.6088	.11484	.013
SPt	196	125.00	255.00	161.4592	27.95956	781.737
DiP	196	.49	.87	.6031	.06904	.005
Valid N (listwise)	196					

Table 2: Attributes description

Attribute	Description	Response
Age	The age of each participant	No
b/a	The 'b' ratio to the 'a' of PPG's second derivative waveform	No
RI	The ration between diastolic peak to systolic peak	No
SPt	Time to reach PPG's systolic peak	No
DiP	Diastolic peak value	No
CIMT	Carotid intima-media thickness	Yes

2. Literature Reviews

Heart diseases are considered killing machines in which they threaten human's life. As they developed silently as we age, early detection and prediction of heart disease is crucial in disease prevention. The ability to accurate diagnosis could lead to better medication and promote better health services. In this study machine learning algorithm namely decision tree is used to classify patients into high-risk atherosclerosis and no-risk atherosclerosis. The proposed classifier, decision tree, classifies the dependent variable named CIMT test based on a training and testing data set contains four independent variables extracted from PPG signal morphology and 'Age' index as a fifth independent variable.

Decision tree is a machine learning algorithm that belongs to a supervised learning algorithms. It normally used for classification problems solving. This study utilized a decision tree classifier to predict the dependent variable based on some derived decision rules extracted from prior data (training and testing phases). It represented as nodes and inter-nodes in which root nodes are used to classify the instances features. While leaf nodes (nodes without children) indicate decisions or classifications. Evaluating highest gain (most homogenous branches) among all other attributes in every stage is the base selection by a decision tree at each node. The performance of a decision tree is evaluated using a confusion matrix.

Sisodia & Sisodia (2018) used three classification algorithms namely: Decision Tree, SVM, and Naive Bayes to detect and predict diabetes. They evaluated the performances of all used algorithms on different measures like accuracy, Recall, and F-Measure. They concluded that Naïve Bayes outperformed decision tree and SVM with an overall accuracy of (76.3%). The role of using the J48 decision tree machine learning methods (Adaboost and Bagging) to classify diabetic patients into diabetic or non-diabetic was discussed by Preveen et al (2016).

Another proposed system by Orabi et al (2016) for diabetes prediction. Their developed system was evaluated using decision tree machine learning. They claimed that their obtained results were satisfactory enough as their system performs fine in predicting diabetes incidents at certain age with high accuracy. Yuvarani & Selvarani (2016) proposed a model for Diabetes prediction using three decision tree classifier which are: LAD tree, Genetic J48 tree, and NB tree. They claimed that this genetic J48 decision tree model has higher prediction accuracy among other used classifier in terms of speed and accuracy. Maji & Arora (2019) proposed a hybridization technique using decision tree and artificial neural network (ANN) classifiers for the prediction of heart disease.

Reddy et al., (2016) proposed a Decision Tree model for the diagnosis of heart disease. They highlighted that using a decision tree classifier in predicting vulnerability of heart disease introduced a reasonable accuracy. In their seeking to develop a smart clinical classifier, Lakshmishree and Paramesha (2017) developed a decision tree based prototype that is able to extract patterns and relations associated with heart diseases. Amin et al., (2018) utilized different machine learning algorithms to classify heart disease patients. The decision tree classifier among other classifiers showed an overall accuracy of 80.4% in classifying subjects into healthy and at high-risk heart disease.

3. Results and Discussions

In this section, the obtained results are discussed. The model was first fit with five independent variables namely b/a, RI, DiP, SPt, extracted from PPG's contour, along with the fifth independent index which is 'Age' index respectively. The model has one dependent variable which is CIMT test score. All variables (independent and dependent) are went through a pre-processing phase in which missing data or outliers are filled or removed. The decision tree classifier then took place in which it checks for the most contributing factors towards CIMT prediction by applying certain rules extracted and learned through training and testing stages. Information gain and best cut-off points were determined through the process of machine learning inside decision tree classifier. The model then predicts and classifies a new input to be high-risk atherosclerosis or no-risk atherosclerosis with an overall accuracy of 82.6%. The diagram of the proposed classification model is shown in figure 1 below.

The decision tree algorithm produced a decision tree classification output in which it is used to predict the response variable. The full classification tree is shown in figure 2. The tree starts by Age index splitting the tree from the root into two main branches, less than 20.5 years or higher than or equals 20.5 years. As shown in figure 2, the left most branch, if your age is less than 20.5 years then definitely the model will classify you as no-risk atherosclerosis (healthy). Because the classification decision comes into two values (0.4 or 0.5) in which both of them are less than the threshold 0.7mm which indicates no risk of atherosclerosis. Tracking the decision tree nodes provides several rules for predicting high-risk atherosclerosis. For example, stating from the root with age index value greater than 20.5 years, moving to the next node with a RI index value less than 0.715, and next node where RI index value is less than 0.625, reaching next node where b/a index is greater than 0.7, continuing with b/a index value greater than 0.77, finally getting the last node in this path where RI index value is greater than or equals to 0.595, the decision tree reached a leaf node which represents a decision in which this particular subject has no-risk atherosclerosis.

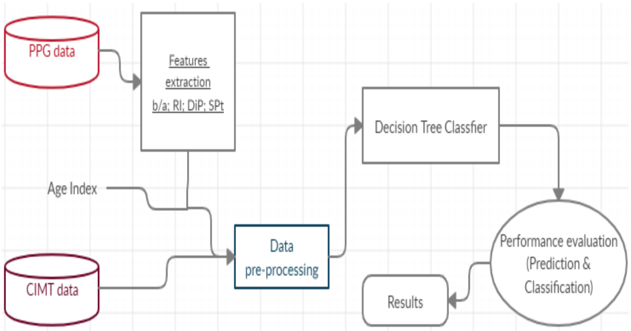

Fig. 1 The proposed classification model diagram

The model relies on certain equation in predicting and classifying the new input data into high-risk atherosclerosis or no-risk atherosclerosis, as described in equation 1 below.

$$newCIMT = predict\left(MNUM\ (newInput)\right) \ldots\ldots\ldots (1)$$

Where, MNUM represent the model name; newInput represents the new data to be fed to the model as described in the following examples.
Examples for predicting and classifying a new data set as high-CIMT value or normal CIMT value. Let us consider the following new input data to the developed model:

(a)

b/a	Age	RI	SPt	DiP
0.61	41	0.72	198	0.8

The model, based on equation 1, will respond as follow:
newCIMT = 0.7100

(b)

b/a	Age	RI	SPt	DiP
0.52	62	0.8	233	0.76

The model, based on equation 1, will respond as follow:
newCIMT = 0.6200

The classification table for predicting high-risk atherosclerosis based on Decision tree classifier is shown in table 3. The model is built on a dataset of 196 subjects. Table 3 introduces the overall percentage of model performance. The model was able to classify 117 patients correctly as no-risk atherosclerosis while it wrongly classified 14 subjects as high-risk atherosclerosis incorrectly with an accuracy (Sensitivity) of 89.3% as described in equation 2. On the other hand, the model showed an ability to classify 45 subjects as having high-risk atherosclerosis while 20 subjects were incorrectly classified as no-risk atherosclerosis with an accuracy (Specificity) of 69.2% as seen in equation 3. However, the model showed a reliable overall accuracy of 82.6% in its ability to predict and classify patients into high-risk or no-risk atherosclerosis (equation 4).

Table 3: Classification Table (prediction accuracy).

		Predicted	
		Negative	Positive
Actual	Negative	(TP)= 117	(FN)=14
	Positive	(FP)= 20	(TN)=45

Based on the results shown in table 4 above, the accuracy, specificity, and sensitivity can be calculated as follow:

$$Sensitivity = \frac{TP}{TP+FN} * 100 = \frac{117}{117+14} * 100 = 89.3\% \ldots\ldots\ldots\ldots\ldots \ldots (2)$$

$$Specificity = \frac{TN}{FP+TN} * 100 = \frac{45}{20+45} * 100 = 69.2\% \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

$$Accuracy = (TP+TN)/\,(TP+FP+TN+FN) * 100 = \frac{117+45}{117+14+45+20} * 100 = 82.6\% \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Another way to examine the performance of the decision tree classifier is to look at the receiver operator characteristic curve (ROC). It represents a visual plot illustrating the diagnostic ability of the model. This technique runs only in binary systems were the output can only be binary (0 or 1; Yes or No; etc.). The true positive rates are plotted against the false negative rates at different cutoff values. Figure 3 demonstrates an ROC curve of b/a index vs. CIMT score. In this example, the b/a index is selected to be the predictor of high-risk atherosclerosis (CIMT). Theoretically, the point (0, 1) represents perfect classification which represents the left-most top edge of ROC curve. As its name implies, theoretically, it is difficult to be achieved in real diagnostic problems. However, the b/a index showed an area under the curve (AUC) of 90.5% which indicates high performance and high accuracy.

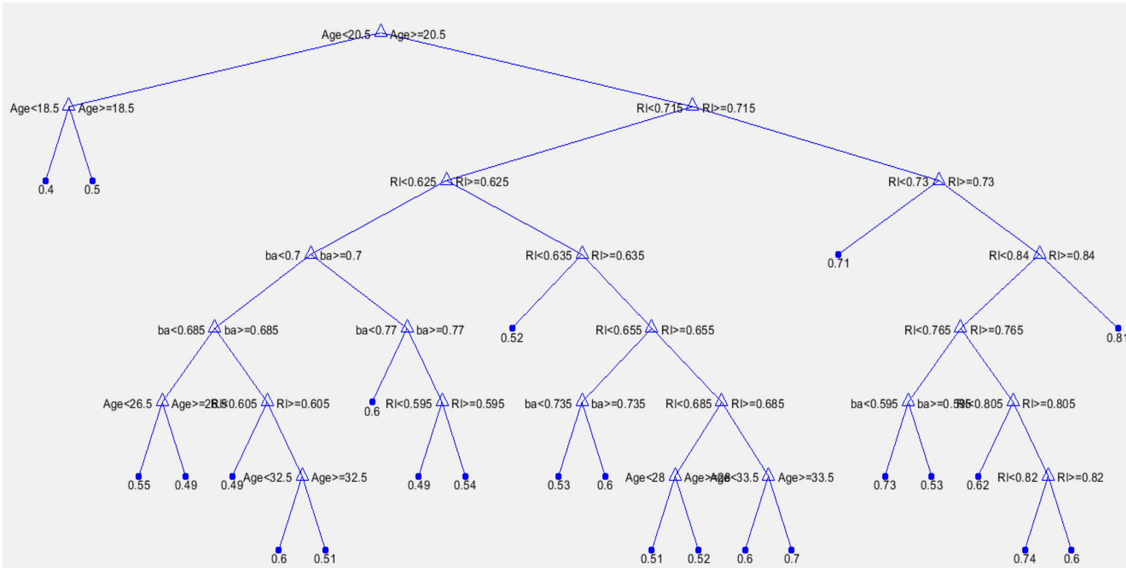

Fig. 2 Decision tree classification of CIMT test.

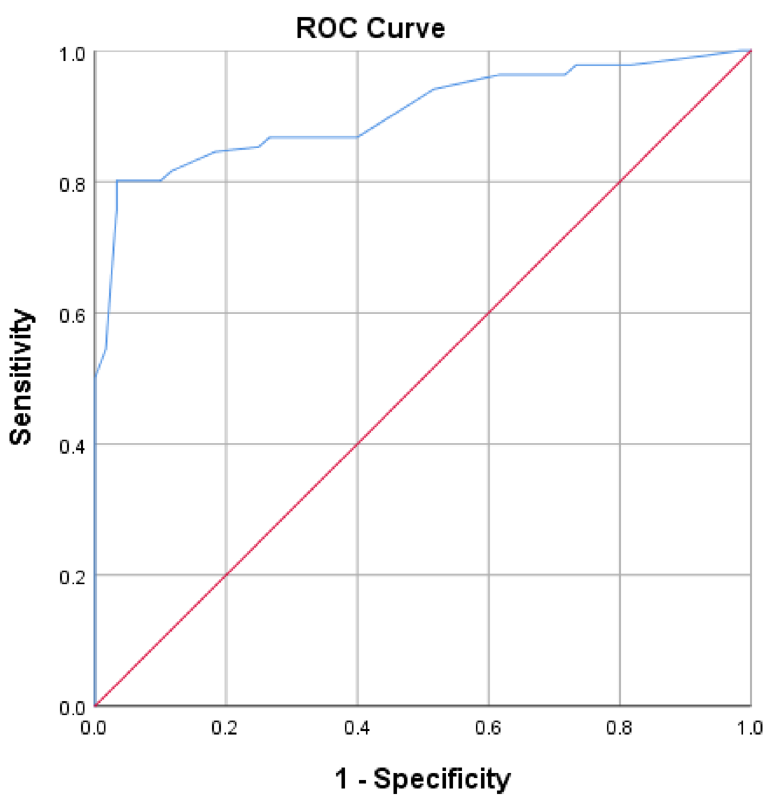

Fig. 3 ROC curve for b/a index in the prediction of CIMT

4. Conclusions

This proposed model is developed for atherosclerosis diagnosis as a surrogate measure to CIMT test. The ability of the developed classifier is measured in terms of overall accuracy, sensitivity, specificity, and the AUC measures. The model obtained an overall accuracy of 82.6%, sensitivity of 89.3%, specificity of 69.2%, and AUC of 90.5% respectively. The results were satisfactory enough since the proposed classifier works fine in predicting the atherosclerosis incidents in which it has high accuracy based on decision tree machine learning classification method. Utilizing features extracted from a non-invasive technique, the PPG, to predict atherosclerosis represent a promising technology that has low cost, mobility, small-size, affordable, and easy setup to be adopted in clinical settings. The proposed decision tree based classifier could assist in atherosclerosis early detection and prevention by predicting its high-risk in an early stage. This without doubt might deliver a fruitful benefit to society and specialists in terms of disease prediction and risk prevention.

References

[1] Qawqzeh Yousef, Uldis Rubins, and Alharbi Mafawez (2015). Photoplethysmogram second derivative review: Analysis and applications. Scientific research and essays 10(21):633-639. DOI: 10.5897/SRE2015.6322

[2] C. Latha & S. Jeeva (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked. Vol 16, pp1-9

[3] Sisodia S & Sisodia D (2018). Prediction of Diabetes using Classification Algorithms. Procedia Computer Science 132 (2018) 1578–1585

[4] Perveen,S., Shahbaz,M., Guergachi,A., Keshavjee,K., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science 82,115–121.doi:10.1016/j.procs.2016.04.016.

[5] Orabi, K. M., Kamal, Y. M.,Rabah, T. M., 2016. Early Predictive System for Diabetes Mellitus Disease. Industrial Conference on Data Mining, Springer. Springer.pp.420–427.

[6] Yuvarani. S and Selvarani. R (2016). AN ANALYSIS OF DECISION TREE MODELS FOR DIABETES. International Research Journal of Engineering and Technology (IRJET). Vol 3, issue 11, pp 680-684

[7] Maji S., Arora S. (2019) Decision Tree Algorithms for Prediction of Heart Disease. In: Fong S., Akashe S., Mahalle P. (eds) Information and Communication Technology for Competitive Strategies. Lecture Notes in Networks and Systems, vol 40. Springer, Singapore.

[8] Reddy. R, Raju. K, Kumar. M, Sujatha. CH, and Prakash. P (2016). Prediction of Heart Disease Using Decision Tree Approach. International Journal of Advanced Research in Computer Science and Software Engineering. Vol 6, issue 3, pp530-532

[9] Lakshmishree. J & Paramesha. K (2017). Prediction of Heart Disease Based on Decision Trees. International Journal for Research in Applied Science & Engineering Technology (IJRASET). Vol 5, issue V, pp943-948

[10] Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun, (2018) “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms,” Mobile Information Systems, vol 2018, pp1-2. https://doi.org/10.1155/2018/3860146.