Intelligent Method for Detection of Coronary Artery Disease with Ensemble Approach

Luxmi Sapra, Jasminder Kaur Sandhu, and Nitin Goyal

Abstract Coronary artery is a major reason of health ailments all over the world. Its detection and management incurred huge amount of deaths across the nation. Heart disease can be diagnosed using various invasive and non-invasive methods. One of the effective methods for detection of coronary artery disease is coronary angiography, which is expensive and also has side effects. This further requires high level of technical expertise. Due to improvement in technology and low-cost storage devices, storage of huge amount of data becomes easy. Even health sector has been untouched. Machine learning methods are being used to analyze the collected data due to its capability to predict the diseases. In this work, machine learning methods are implemented in order to achieve low-cost, reproducible, non-invasive, rapid, and precise identification of heart disease. This paper adopted ensemble method with multiple classifiers to construct and validate the model. For experiment purpose, Z-Alizadesh Sani coronary artery disease dataset is used. The ensemble method of prediction outperforms the other disease prediction methods.

Keywords Ensemble method · Coronary artery disease · Non-invasive · Angiography

1   Introduction

Cardiovascular diseases are one of the foremost reasons of death and disability all over the country. Deaths due to these diseases are increasing day by day in every age group. These are basically the disease of heart and blood vessels. Coronary artery disease is a category of cardiovascular disease where the plaque is accumulated in heart vessels, interrupts the flow of blood, causes pain, and results into heart attack or even death [1–4]. The major causes of heart diseases are unhealthy lifestyle, lack of exercise, smoking, and unhealthy eating habits. Developing as well as developed

L. Sapra · J. K. Sandhu · N. Goyal (✉)
Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India
e-mail: nitin.goyal@chitkara.edu.in

countries are spending large amount of its nation financial budget for detection and treatment of the disease. In spite of advancement in medical science, accurate diagnosis and treatment of the heart disease is still a challenging task because of the complexity of diagnosis and treatment methods, especially in resource poor settings. Accurate diagnosis and treatment of patients is necessary in order to save human life and also to reduce the risk of more severe disease [5–9]. Heart disease is prevented by adopting healthy life style and timely tracking off and treatment of the disease. There are many well-known invasive and non-invasive modalities available for identification of the disease [10]. Non-invasive methods include techniques such as ECG, echocardiogram, stress test, and so on. Sometimes the result of these modalities is inconclusive as well as requires time for assessment. So, coronary angiography is popular examining modality for disease identification. It is invasive, painful, and expensive as well as requires expensive clinical setup [11–14].

Due to improvement in technology and low-cost storage devices, storage of huge amount of data becomes easy. Even health sector has been untouched. Machine learning methods are being widely used to analyze the collected data due to its capability to predict the diseases. Researchers are seeking inexpensive, reproducible, fast, and computationally inexpensive methods for detection of heart disease. They are exploring machine learning methods such as support vector machine (SVM), K-nearest neighbor (KNN), artificial neural network (ANN), decision tree, logistic regression, and naive Bayes for identification of heart disease [15–20]. The paper focuses on various machine learning methodologies in order to identify the heart disease. For experiment purpose, Cleveland heart disease dataset and Alizeshani heart disease dataset available at University of California Irvine (UCI) machine learning repository are used.

Nowadays, healthcare sector is generating large amount of data related to patients, disease, clinical reports, physician notes, laboratory tests, and administrative data. The collected data is used by knowledge miners to extract useful patterns with the help of advanced computational intelligent methods results in low-cost healthcare services with reduced error, improved diagnostic methods.

2   Framework for Intelligent Coronary Artery Disease Prediction

The benchmark coronary artery disease dataset is collected from UCI machine learning repository available for research purpose. The Z-Alizadesh Sani coronary artery dataset contains 53 attributes and 303 records such as Age, Weight, Length, Gender, Body mass index, Diabetes Mellitus, Hypertension, Current smoker, EX-Smoker, Obesity, Airway disease, Thyroid, Chest pain, etc. The Cleveland data consists of 14 attributes and 303 instances. The features are Age, Gender, Chest pain type, Resting blood pressure on admission, Serum cholesterol, Fasting blood

sugar, Resting ECG outcome, Max heart rate achieved, Old peak, Slope, Number of fluoroscopy, Colored vessels, Reversible defect, and Outcome [21].

Data is preprocessed to apply predictive modeling using classification methods. The disease identification model is evaluated using performance measures such as accuracy, error rate, AUC, and F-measures. The experimental results exhibit that ensemble-based model which is a better approach with regard to reliability and predictivity of diagnosis. Table 1 shows the description of Z-Alizadesh Sani heart disease dataset.

The CAD dataset collected from UCI machine learning repository was preprocessed. Then, logistic regression, deep learning, decision tree, random forest, gradient boosted, and SVM learning algorithm were applied to identify the presence and absence of coronary artery disease. The performance measures to evaluate the recital of learning algorithms are recorded such as accuracy, error rate, AUC, and F-measure.

Accuracy parameter outputs the percentage of correctly identified patients keeping into consideration following the observations of people suffering from coronary artery disease and those who are not suffering from this disease. Error rate is the percentage of patients not suffering from coronary artery disease, identified as positive for disease, and patients who are suffering from disease and identified as negative for the diseases. Figure 1 shows the framework for intelligent coronary artery disease system.

3   Result Analysis

Table 2 presents the result on Z-Alizadesh Sani dataset having accuracy, error rate, F-measures, and AUC. Logistic regression achieves the prediction accuracy of 80%, deep learning-based model achieves the prediction accuracy of 78%, decision tree achieves the prediction accuracy of 69%, SVM achieves accuracy of 71%, and random forest achieves the prediction accuracy of 82%. The Ensemble-based model gradient boosted tree achieves highest prediction accuracy of 84%. Figure 2 presents the accuracy of classification methods.

In case of misclassification error rate, logistic regression achieves the error rate 20%, deep learning-based model achieves the error rate of 22%, decision tree achieves the error rate of 31%, random forest achieves the error rate of 18%, SVM achieves the error rate of 29%, and in case of gradient boosted tree, it achieves the lowest error rate of 16%. Figure 3 presents the misclassification error rate of classifiers. Figures 4 and 5 present the AUC and F-measure of prediction models.

Table 3 shows the experimental results on Cleveland heart disease dataset. Logistic regression achieves the prediction accuracy of 83%, deep learning achieves the prediction accuracy of 83%, and decision tree achieves the prediction accuracy of 74%. Random forest achieves the prediction accuracy of 77%, SVM achieves the prediction accuracy of 74%, and gradient boosted tree (the ensemble-based method) achieves the highest prediction accuracy of 84%. On the other hand, gradient boosted tree achieves the lowest error rate of 16%, and support vector machine has the highest

Table 1 Description of Z-Alizadesh Sani dataset

Attribute name	Type	Min	Max
Age	Integer	30.0	86.0
Weight	Integer	48.0	120
Length	Integer	140.0	188.0
Sex	Categorical	Male (176)	Female (127)
Body mass index	Real	18.11	40.9
Diabetes Mellitus	Integer	0.0	1.0
Hypertension	Integer	0.0	1.0
Current smoker	Integer	0.0	1.0
EX-smoker	Integer	0.0	1.0
Family history	Integer	0.0	1.0
Obesity	Categorical	Y (211)	N (92)
Chronic renal failure	Categorical	N (297)	Y(6)
Cerebrovascular accident	Categorical	N (298)	Y(5)
Airway disease	Categorical	N (292)	Y(11)
Thyroid disease	Categorical	N (296)	Y(7)
Congestive heart failure	Categorical	N (302)	Y(1)
Dyslipidemia	Categorical	N (191)	Y(112)
Blood pressure	Integer	90.0	190.0
Pulse rate	Integer	50.0	110.0
Edema	Integer	0.0	1.0
Weak peripheral pulse	Categorical	N (298)	Y(5)
Lung rales	Categorical	N (292)	Y(11)
Systolic murmur	Categorical	N (262)	Y(41)
Diastolic Murmur	Categorical	N (294)	Y(9)
Typical chest pain	Categorical	0.0	1.0
Dyspnea	Categorical	N (169)	Y (134)
Function class	Integer	0.0	3.0
Atypical	Categorical	N (210)	Y (93)
Nonanginal	Categorical	N (287)	Y(16)
Exertional chest pain	Categorical	N (303)	NA
Low threshold angina	Categorical	N (301)	Y{2}
Q wave	Integer	0.0	1.0
St elevation	Integer	0.0	1.0
St depression	Integer	0.0	1.0
Tinversion	Integer	0.0	1.0
left ventricular hypertrophy	Categorical	N (283)	Y(20)

(continued)

Table 1 (continued)

| Attribute name	Type	Min	Max
Poor R wave regression	Categorical	N (294)	Y(9
Bundle branch blocks	Categorical	N (282)	LBBB (13)
Fasting blood sugar	Integer	62.0	400.0
Creatine	Real	0.5	2.2
Triglyceride	Integer	37.0	1050.0
Low density lipoprotein	Integer	18.0	232.0
High density lipoprotein	Integer	15.0	111.0
Blood urea nitrogen	Integer	6.0	52.0
Erythrocyte sedimentation	Integer	1.0	90.0
Hemoglobin	Real	8.9	17.6
Potassium	Real	3.0	6.6
Sodium	Integer	128.0	156.0
White blood Cells	Integer	3700.0	18000.0
Lyniph	Integer	7.0	60.0
Neutrophil	Integer	32.0	89.0
Platelet	Integer	25.0	742.0
Ejection fraction	Integer	15.0	60.0
Regional wall motion abnormality	Integer	0.0	4.0
Valvular heart disease	Categorical	Mild (149)	N (116)
CAD	Categorical	Yes (216)	N (87)
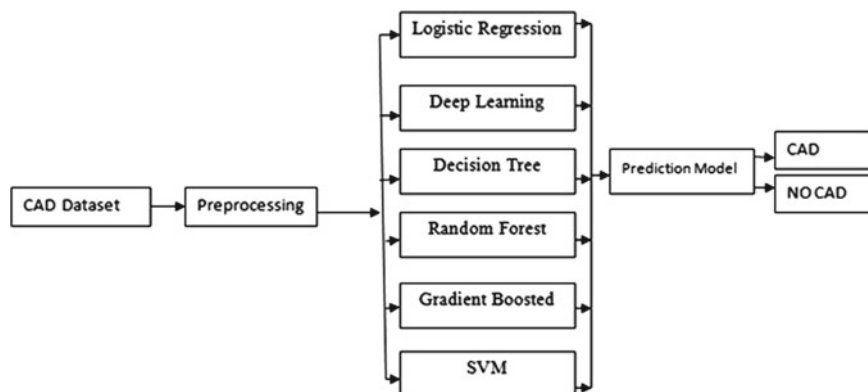
Fig. 1 Framework for intelligent coronary artery disease system

Table 2 Accuracy/Error rate/F-measures/AUC of models on Z-Alizadesh Sani dataset

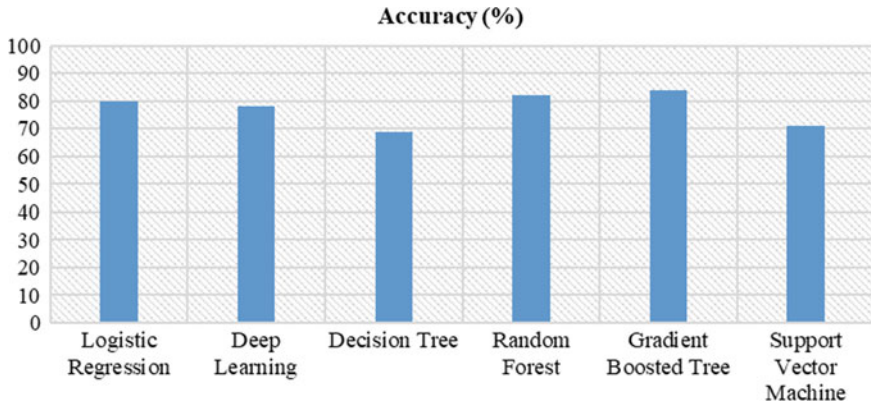| Model	Accuracy (%)	Error rate (%)	F-measure	AUC
Logistic regression	80	20	85	0.551
Deep learning	78	22	84	0.888
Decision tree	69	31	81	0.499
Random forest	82	18	88	0.923
Gradient boosted tree	84	16	89	0.881
Support vector machine	71	29	83	0.508

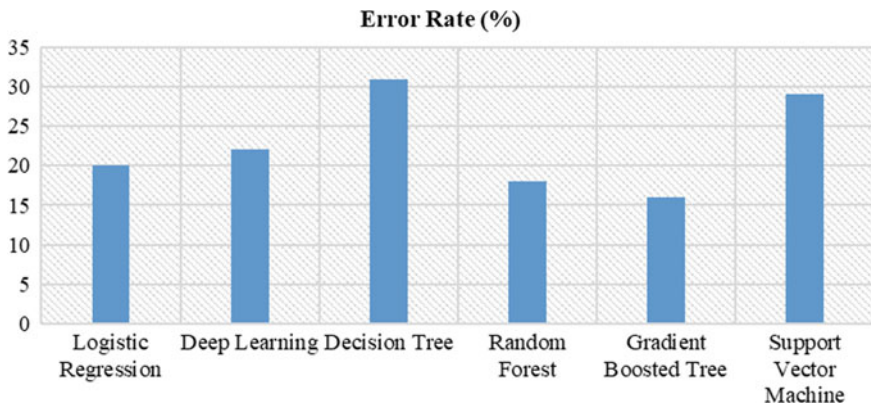**Fig. 2** Accuracy of models on Z-Alizadesh Sani dataset

Fig. 3 Error rate of models on Z-Alizadesh Sani dataset
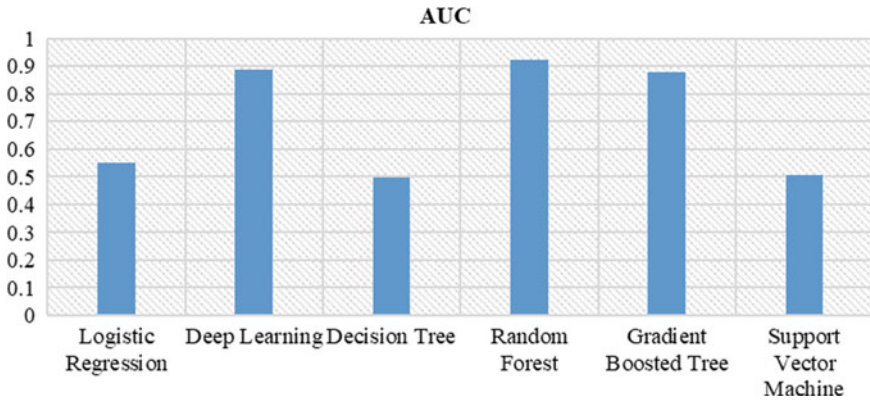
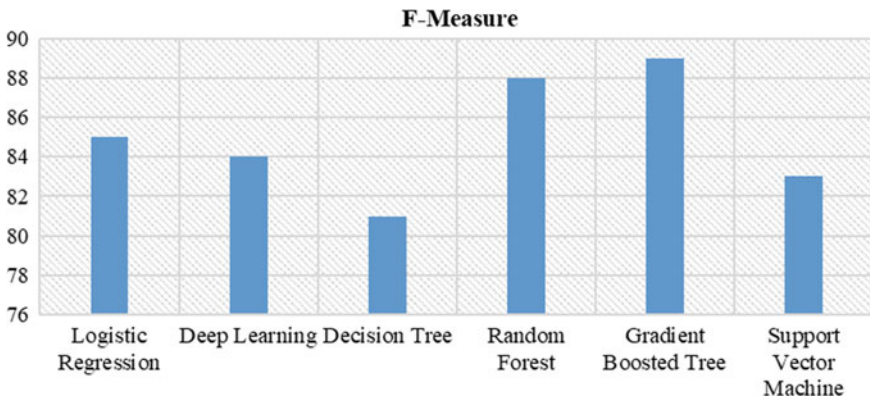**Fig. 4** AUC of models using Z-Alizadesh Sani dataset

Fig. 5 F-measure of models using Z-Alizadesh Sani dataset

Table 3 Results on Cleveland heart disease dataset

Model	Accuracy (%)	Error rate (%)	AUC
Logistic regression	83	17	0.905
Deep learning	83	17	0.888
Decision tree	74	26	0.782
Random forest	77	23	0.893
Gradient boosted tree	84	16	0.905
Support vector machine	74	26	0.82

error rate of 26%. Logistic regression achieves the error rate of 17%, deep learning 17%, random forest 23%, respectively, (Figs. 6, 7 and 8).
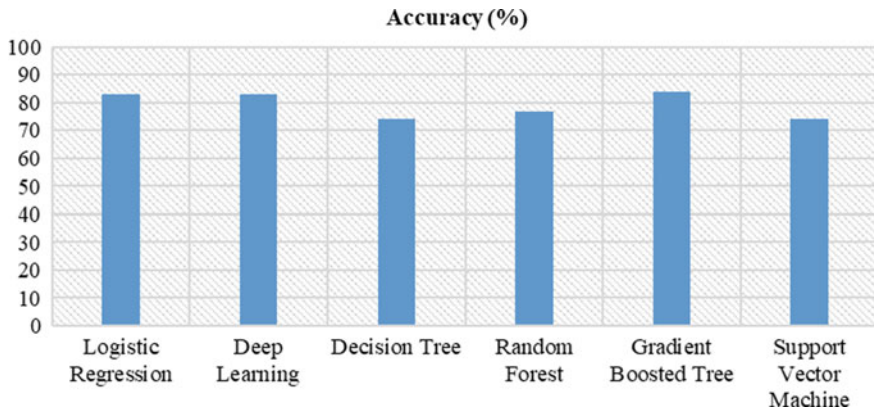
Accuracy (%)

Fig. 6 Accuracy of models using Cleveland heart disease dataset

Error Rate (%)

Fig. 7 Error rate of models using Cleveland heart disease dataset
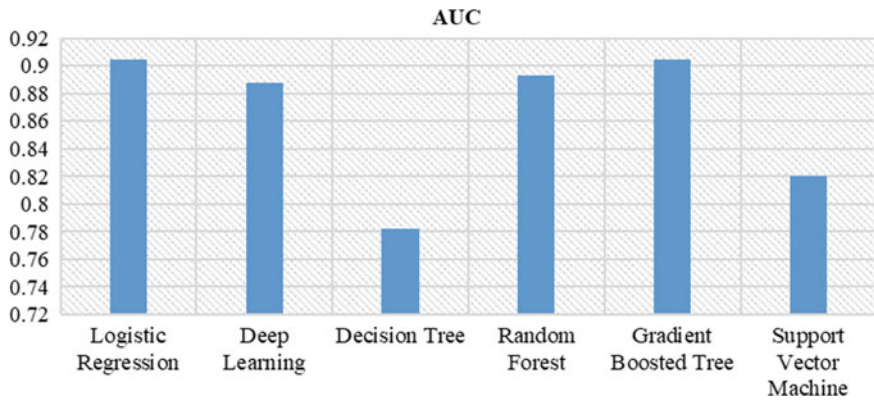
AUC

Fig. 8 AUC of models using Cleveland heart disease dataset

4 Conclusion

The experimental results using Z-Alizadesh Sani dataset and Cleveland datasets show that ensemble-based method is preferred as compared to other coronary artery disease detection models. The proposed model can be used to reduce cost of initial detection of coronary artery disease with low cost. The clinical parameters can be easily collected from hospitals, and results can be reproduced in a faster, more accurate, scalable, and reliable manner. It can serve an adjunct tool in clinical settings.

References

1. http://www.who.int/mediacentre/factsheets/fs317/en/
2. Terzic A, Waldman S (2011) Chronic diseases: the emerging pandemic. Clin Transl Sci 4(3):225–226. https://doi.org/10.1111/j.1752-8062.2011.00295.x
3. Reddy KS, Shah B, Varghese C, Ramadoss A (2005) Responding to the threat of chronic diseases in India. The Lancet 366(9498):1744–1749
4. Verma L, Srivastava S, Negi PC (2018) An intelligent noninvasive model for coronary artery disease detection. Complex Intell Syst 4(1):11–18
5. Wong ND (2014) Epidemiological studies of CHD and the evolution of preventive cardiology. Nat Rev Cardiol 11(5):276–289
6. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM (2018) Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLoS ONE 13(8):e0202344
7. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP (2018) Machine learning in cardiovascular medicine: are we there yet? Heart 104(14):1156–1164
8. Verma L, Srivastava S (2016) A data mining model for coronary artery disease detection using non-invasive clinical parameters. Indian J Sci Technol 9(48):1–6
9. Tiwaskar SA, Gosavi R, Dubey R, Jadhav S, Iyer K (2018) Comparison of prediction models for heart failure risk: a clinical perspective. In: 2018 Fourth international conference on computing communication control and automation (ICCUBEA). IEEE, pp 1–6
10. Verma L, Srivastava S, Negi PC (2016) A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. J Med Syst 40(7):1–7
11. Takahashi N et al (2019) Computerized identification of early ischemic changes in acute stroke in noncontrast CT using deep learning. In: Medical imaging 2019: computer-aided diagnosis, vol 10950. International Society for Optics and Photonics
12. Alizadehsani R, Hosseini MJ, Sani ZA, Ghandeharioun A, Boghrati R (2012) Diagnosis of coronary artery disease using costsensitive algorithms. In: 2012 IEEE 12th international conference on data mining workshops (ICDMW). IEEE, pp 9–16
13. Arafat S, Dohrmann M, Skubic M (2005) Classification of coronary artery disease stress ECGs using uncertainty modeling. In: 2005 ICSC congress on computational intelligence methods and applications. IEEE, pp 4-pp)
14. Sandhu JK, Verma AK, Rana PS (2018) A data-driven framework for survivable wireless sensor networks. In: 2018 Eleventh international conference on contemporary computing (IC3). IEEE, pp 1–6
15. Ayatollahi H, Gholamhosseini L, Salehi M (2019) Predicting coronary artery disease: a comparison between two data mining algorithms. BMC Public Health 19(1):448
16. Dhanaseelan R, Sutha MJ (2018) Diagnosis of coronary artery disease using an efficient hash table based closed frequent itemsets mining. Med Biol Eng Comput 56(5):749–759

17. Alizadehsani R, Habibi J, Sani ZA, Mashayekhi H, Boghrati R, Ghandeharioun A, Bahadorian B (2012) Diagnosis of coronary artery disease using data mining based on lab data and echo features. J Med Bioeng 1(1)
18. Bouali H, Akaichi J (2014) Comparative study of different classification techniques: heart disease use case. In: 2014 13th International conference on machine learning and applications (ICMLA). IEEE, pp 482–486
19. Sandhu JK, Verma AK, Rana PS (2018) A novel framework for reliable network prediction of small-scale wireless sensor networks (SSWSNs). Fundamenta Informaticae 160(3):303–341
20. Acharya UR, Faust O, Sree V, Swapna G, Martis RJ, Kadri NA, Suri JS (2014) Linear and nonlinear analysis of normal and CAD-affected heart rate signals. Comput Methods Prog Biomed 113(1):55–68
21. https://archive.ics.uci.edu/ml/datasets/Heart+Disease