Advances in Parallel Computing Technologies and Applications
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC210160

Heart Disease Prediction Using Hybrid Random Forest Model Integrated with Linear Model

Jaishri Pandhari Wankhede [a,1], Palaniappan S [b] and Magesh Kumar S [a]
[a] *Department of CSE, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, SIMATS*
[b] *Department of CSE, KCG College of Technology, Anna University*

Abstract. The objective of the paper is to throw light on few existing heart disease predicting approaches and proposes a Hybrid Random Forest Model Integrated with Linear Model (HRFMILM) for predicting and identifying the HDs at an early stage. Even though the linear model has simple estimation procedure, it is very sensitive to outliers and may lead to overfitting process. On the other hand, averaging in Random Forest Model (RFM) improves the overall accuracy and reduces the possibility of overfitting. The dataset is collected from standard UCI repository. Experimental results concluded that the integration of Linear Model with RFM makes the simple estimation procedure with improved overall accuracy than the respective models. Further, the proposed method compares the prediction performance of few existing approaches in terms of parameters, namely, precision, recall and F1-score.

Keywords. Heart Disease, Linear Model, Random Forest Model, Hybrid Model, prediction parameters.

1. Introduction

The intervention of a cardiologist arises if a patient suffers from all or any of the following symptoms - shortness in breathing, chest discomfort, nausea and fatigue or any of this kind. The prevailing medical society utters heart disease to be an essential reason for death. Without the discrimination of age or gender, cardiovascular disease crop up for many reasons such as family history, smoking, high blood pressure and cholesterol levels, obesity, poor hygiene and stress [1, 2]. Further, a thorough medical history of the patient's family and the patient plays a vital in this discussion. Based on the reports, the doctor decides either initiate the treatment or to perform the invasive test (coronary cardiography, catheterization, electrophysiology study) [3–5]. Heart diseases are categorized as congenital, arrhythmia, coronary artery, dilated cardiomyopathy, hypertrophic cardiomyopathy, heart failure, pulmonary stenosis, mitral reurgitation based on the organs they affect [1]. The symptoms of heart disease vary for elderly and diabetic patients. Henceforth, an MRI reduces the dimensionality of the heart disease by depicting

[1] Jaishri Pandhari Wankhede, Department of CSE, Saveetha Institute of Medical & Technical Sciences, India. E-mail: wankhede.jaishri@yahoo.com.

the structure of the heart including valves, muscles, chambers other vessels and analyzes the blood flow, narrow or blocked arteries.

Nevertheless, the disease, when recognized earlier, makes the treatment unconstrained forasmuch as identification, a challenging task depends solely on the medical community. The affordability to treat such disease needs to be in the limits of the patient. Data mining is modest as data is readily available. Figure 1 depicts the conventional data mining architecture. The responsibility of pattern evaluation module is to measure the patterns based on the threshold value and interacts with users and data mining engine. The knowledge base provides input to data mining engine for tuning the performance of data mining process.

2. Decision Support Systems

Decision making is an intellectual task with complexity. Using internet and other web tools people acquire knowledge to take smart decisions. Moreover, the list of mentor – friend, colleague is expanded as subject specific. Decision Support Systems are computer based systems intended to help in effective decision making through the data and the models. A decision support system has three integrants i) knowledge ii) inference engine iii) user support. The medical domain has cuddled new information and communication technology through data mining and DSS, to provide economically feasible quality healthcare services for the patients in demand.

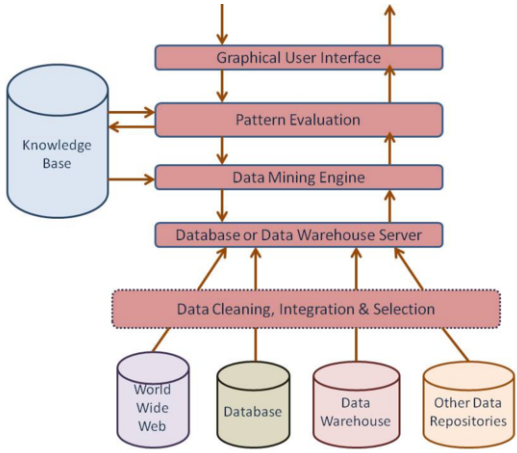

Figure 1. Data Mining Architecture

Foremost, the knowledge in DSS though typified in four forms - logical conditions, rules, graphs and structures, with respect to DSS in diagnosis and treatment, Guideline Interchange Format (GLIF), Clinical Terminology System, SNOMED CT (Systematized Nomenclature of Medicine Clinical Terminology) can be utilized.

The following section illustrates few existing DSS that were adopted for HD predictions and the proposed HD prediction approach.

3. Literature Survey

Vennemann et al. [6] used detection algorithms along with machine learning to diagnose the heart valve degradation based on blood flow data in the heart. They acclaim the diagnosis process as automatic that reduced the frequent visits to clinical checkups and utilized computer-oriented diagnostics and telemedicine for improved patient care. The novelty detection algorithm exhibited high sensitivity of Model Compliance Index (MCI) with respect to Aortic regurgitation, a common heart valve disease where the valve doesn't close completely. The datamining approach for disease prediction provides flexiblity and availability of symptoms checker for wide variety of audience [7]. Thomford et al. [8] reviewed various heart prediction papers that used techniques Artificial Intelligence and digital health system with limited resources settings especially with respect to Sub-Saharan African (SSA) countries. Swapna et al. [9] detected cardiac arrhythmia through automatic detection technique using deep learning. The abnormality of heart beats, cardiac arrhythmia was detected with techniques such as Convolutional Neural Network (CNN), recurrent neural network (RNN), gated recurrent unit, long short-term memory (LSTM) and hybrid of CNN and recurrent structures.

Alarsan and Younes [10] used machine learning algorithms such as Decision tree, Random Forest (RF) and Gradient Boost algorithms (GDB) to classify and analyze heart diseases based on ECG signals. ECG signals of MIT-BIH arrhythmia and MIT-BIH Supra ventricular arrhythmia were processed using Spark-Scala tools to extract features. Choi et al. [11] evaluated the diagnostic accuracy of Artificial intelligence based Clinical decision support system in diagnosing heart failures. The pilot study with a dataset of 97 patients investigated based on knowledge acquisitions - expert driven, data driven also a hybrid of both.

Nashif et al. [12] diagnosed heart diseases utilizing machine learning algorithms with an exclusive health monitoring system for cardiac diseases. The machine learning algorithm, Support Vector Machine (SVM) showed 97.53% in 10 fold cross validation on Cleveland Heart Disease dataset with 303 records along with Statlog dataset with 270 records having 13 similar features. Latha and Jeeva [13] applied ensemble classification technique to improve the accuracy of heart disease prediction. Ensemble classification which uses multiple algorithms overcomes the prediction accuracy of weak classifiers in addition to disease prediction at infancy state itself.

Sharanyaa et al. [14] diagnosed the heart diseases by applying Machine Learning (ML) algorithms such as K-Nearest Neighbor (KNN), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM) and finally a hybrid of all afore mentioned techniques on a public dataset repository, UCI. The investigation showed hybrid of ML techniques exhibited 94% accuracy in heart disease prediction. Jackins et al. [15] depicted smart clinical disease prediction through ML techniques Naïve Bayes (NB) and Random Forest (RF) classifiers. The study was examined on three different datasets with clinical disease such as diabetes, coronary heart disease and breast cancer. Also study by Jabbar et al. [16] on heart disease prediction system using random forest produced good results. Alotaibi [17] implemented a machine learning model to predict heart failure disease accurately. The investigation was conducted on Cleveland dataset with 14 attributes. The ML algorithms considered for the study are NB, DT, SVM, Logistic Regression, and RF. The class precision and class recall for the ML algorithms as confusion matrix form were tabulated and compared. Comparative results depicted DT outperformed other four models.

4. Proposed Approach

The proposed work integrates the Linear Model (LM) and Random Forest (RF) to form a hybrid approach, namely Hybrid Random Forest Integrated with Linear Model (HRFMILM) for predicting and identifying the Heart Diseases at an early stage. The Cleveland HD dataset is downloaded from UCI repository for evaluating the proposed work. The dataset comprises of 76 attributes collected from 303 individuals. Various attributes collected from the individuals considered in the proposed method are (1) Age (2) Sex (3) Type of the chest pain (4) Blood pressure at rest (5) Serum cholesterol (6) Blood sugar at fasting (7) ECG at rest (8) Maximum Heart Rate (HR) (9) Exercise induced during angina (10) ST depression created due to exercise (11) ST level at peak exercise (12) Major vessels count (13) Thal and (14) HD diagnosed result.

The following section illustrates the steps for HD prediction using Linear Model.

Step 1 : Input the dataset
Step 2 : Handling the missing values
Step 3 : Splitting the dataset for training and testing purposes. In this approach, 75% and 25% ratio of dataset is adopted for training and testing processes respectively.
Step 4 : Feature Scaling is carried out for standardizing the independent features in the dataset for training and testing samples.
Step 5 : Fitting the LM classifier (in this case, the K-Nearest Neighbour (KNN) with neighbour (n)=3)
Step 6 : Predicting the classifier performance

The following section illustrates the steps for HD prediction using RFM approach.

Step 1 : Import the dataset
Step 2 : Handling the missing values
Step 3 : Splitting the dataset for training and testing purposes. In this approach, 75% and 25% ratio of dataset is adopted for training and testing processes respectively.
Step 4 : Feature Scaling is carried out for standardizing the independent features in the dataset for training and testing samples.
Step 5 : Fitting the RFM classifier with the number of estimator equals to twenty.
Step 6 : Predicting the classifier performance.

5. Results And Discussion

The proposed method is evaluated on the Cleveland HD dataset downloaded from UCI repository. The various parameters used to evaluate the proposed method are Precision, Recall, F1-score and accuracy. The Precision (P), Recall (r), F1-score and Accuracy (A) are determined using the Eqs. (1), (2), (3) and (4), respectively.

$$Precision(P) = \frac{TP}{(TP+FP)} \tag{1}$$

$$Recall(r) = \frac{TP}{(TP+FN)} \tag{2}$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

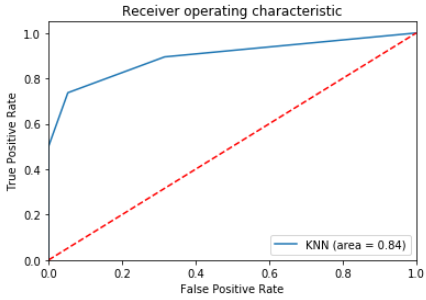

Figure 2. ROC for LM (KNN)

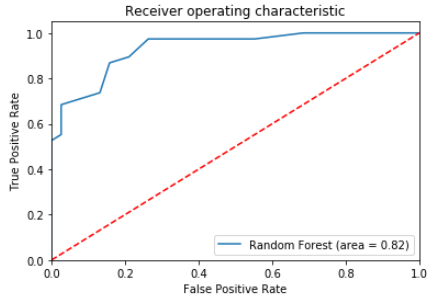

Figure 3. ROC for RFM

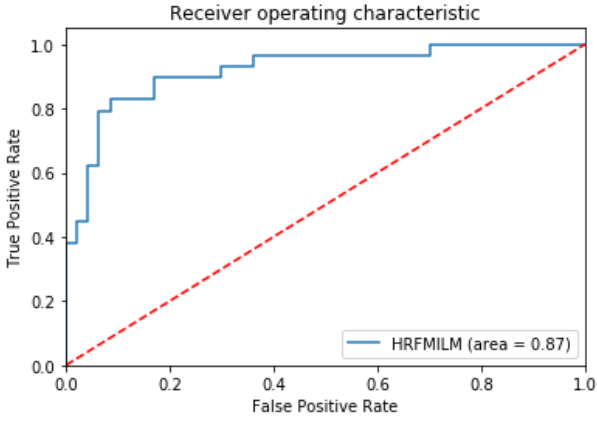

Figure 4. ROC for proposed HRFMILM

Figure 2, 3 and 4 illustrate the performance of LM, RFM and proposed HRFMILM respectively regarding the Receiver Operating Characteristics (ROC) curve. The ROC curve is defined as the plot drawn between the two parameters, namely, True Positive Rate (TPR) and False Positive Rate (FPR). TPR is also referred as recall (r) and FPR is defined using the Eq. (5). Table 1 illustrates the performance comparison of LM, RFM and proposed HRFMILM.

$$FPR = \frac{FP}{(FP + TN)} \tag{5}$$

Table 1. Performance comparison

Name of the Method	Precision (P)	Recall (r)	F1-score	Accuracy in %
LM	0.85	0.85	0.84	84
RFM	0.83	0.82	0.81	82
Proposed HRFMILM	0.83	0.82	0.81	87

6. Conclusion

To summarize, the paper presented a HD prediction approach using Hybrid Random Forest Model Integrated with Linear Model (HRFMILM). The present work is evaluated on Cleveland dataset collected from UCI repository. When compared to the performance of the conventional machine learning approaches regarding HD prediction, the integration of those traditional approaches resulted in improved and reliable HD prediction accuracy. Henceforth, the proposed approach has probable of diagnosing the HD to aid cardiologists. As a part of future work, proposed work can be integrated with few other traditional machine learning approaches and evaluated for performances and even with other datasets.

References

[1] Peter TJ, Somasundaram K. An empirical study on prediction of heart disease using classification data mining techniques. InIEEE-International conference on advances in engineering, science and management (ICAESM-2012) 2012 Mar 30 (pp. 514-518). IEEE.

[2] Everything you need to know about heart disease, - Heart Disease: Types, Causes, And Treatments [Internet]. 2020 Sep 03 [cited 2020 Dec 31]. Available from: https://www.medicalnewstoday.com/ articles/237191.

[3] Mayo Clinic. Heart Disease - Symptoms And Causes [Internet]. 2021 Jan 12 [cited 2020 Dec 30]. Available from: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118.

[4] Magnetic Resonance Imaging (MRI) [Internet]. [cited 2020 Dec 30]. Available from: https://www.heart.org/en/health-topics/heart-attack/diagnosing-a-heart-attack/magnetic-resonance-imaging-mri.

[5] Booma PM, Prabhakaran S, Dhanalakshmi R. An Improved Pearson's Correlation Proximity-Based Hierarchical Clustering for Mining Biological Association between Genes. The Scientific World Journal. 2014 Jan 1;2014:1-10.

[6] Vennemann B, Obrist D, Rösgen T. Automated diagnosis of heart valve degradation using novelty detection algorithms and machine learning. PloS one. 2019 Sep 26;14(9):e0222983.

[7] Wideskills. Data Mining Architecture, Data Mining Tutorial [Internet]. [cited 2020 Dec 31]. Available from: https://www.wideskills.com/data-mining-tutorial/data-mining-architecture.

[8] Thomford NE, Bope CD, Agamah FE, Dzobo K, Owusu Ateko R, Chimusa E, Mazandu GK, Ntumba SB, Dandara C, Wonkam A. Implementing artificial intelligence and digital health in resource-limited settings? Top 10 lessons we learned in congenital heart defects and cardiology. Omics: a journal of integrative biology. 2020 May 1;24(5):264-77.

[9] Swapna G, Soman KP, Vinayakumar R. Automated detection of cardiac arrhythmia using deep learning techniques. Procedia computer science. 2018 Jan 1;132:1192-201.

[10] Alarsan FI, Younes M. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. Journal of Big Data. 2019 Dec;6(1):1-15.

[11] Choi DJ, Park JJ, Ali T, Lee S. Artificial intelligence for the diagnosis of heart failure. NPJ digital medicine. 2020 Apr 8;3(1):1-6.

[12] Nashif S, Raihan MR, Islam MR, Imam MH. Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. World Journal of Engineering and Technology. 2018 Sep 12;6(4):854-73.

[13] Latha CB, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked. 2019 Jan 1;16(100203):1-9.
[14] Sharanyaa S, Lavanya S, Chandhini MR, Bharathi R, Madhulekha K. Hybrid Machine Learning Techniques for Heart Disease Prediction. International Journal of Advanced Engineering Research and Science. 2020;7(3):44-8.
[15] Jackins V, Vimal S, Kaliappan M, Lee MY. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing. 2021 May;77(5):5198-219.
[16] Jabbar MA, Deekshatulu BL, Chandra P. Intelligent heart disease prediction system using random forest and evolutionary approach. Journal of network and innovative computing. 2016 Apr;4(2016):175-84.
[17] Alotaibi FS. Implementation of machine learning model to predict heart failure disease. International Journal of Advanced Computer Science and Applications. 2019;10(6):261-8.