Computational Intelligence for Detection of Coronary Artery Disease with Optimized Features

Varun Sapra, Madan Lal Saini

Abstract*: *Coronary Artery Disease (CAD) is one of the foremost cause of mortality in almost all over the world. It falls under the category of non-communicable diseases, that are spreading at a faster pace nowadays. The factors that create a domino effect on the disease are changing life styles, unhealthy food habits, lack of exercise and other socioeconomic factors. In the past few years, with the advancement in information technology services, health sector is transformed largely and is transmitting a massive amount of medical information. With the advancement of machine learning intelligent computational methods have proved their effectiveness in almost every field. Medical field is also getting benefitted from machine learning because of its capabilities to model complex relations. This paper discusses the use of Firefly for feature subset selection with different machine learning schemes for the identification of CAD. The different techniques implemented are Random Forest, Fuzzy Unordered Rule Induction, Logistic regression and Multilayer perceptron using Keras. Deep learning based method outperforms other learning schemes with the accuracy of 89.77%. Thus, the method can pose as a promising tool for screening CAD patients more accurately.

Index Terms*: *Cardiovascular Disease, Coronary Artery Disease, Feature Subset selection, Multilayer Perceptron

I. INTRODUCTION

Coronary Artery Disease (CAD) is considered as one of the chronic illness that is growing at 9.2% annually, and by 2030 cardio vascular diseases will be the leading cause of deaths all over the world accounting for almost one third of deaths. Health sector is facing a major challenge for handling non-communicable diseases like CVD, as by 2030 seven out of ten diseases will be non-communicable disease [1].

Health sector is one of the growing sector in all the countries and with the growth of technology, which is a palpable reason; this sector is generating a huge amount of complex medical data about patients, prescriptions, medical infrastructure and disease diagnosis. This huge amount of data contains complex relationships, and hospitals needs to increase their ability to analyze and formulate better ways of retrieving these hidden patterns to support clinical decision making. The major reason for CAD is the disordering of heart and narrowing of blood vessels due to the presence of plaque in arteries. The plaque consists of extracellular matrix, which consists of cholesteryl esters, T lymphocytes, cholesterol, phospholipids and various connective tissues.

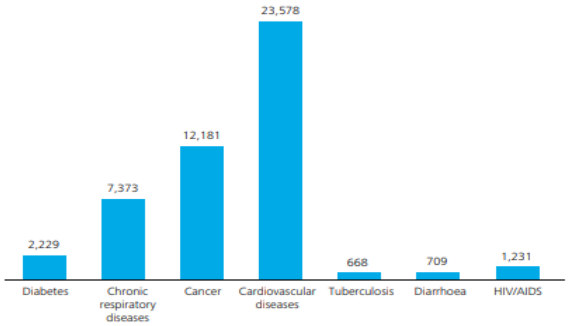

Figure 1: Major causes of mortality by 2030 [1]

It restricts or reduces the flow of oxygenated blood to the heart muscle [2-5]. Coronary angiography is the most prominent method for detection of coronary artery disease. Due to limitation of Coronary angiography like high cost, the complex, and painful carrying out method, encouraged researchers to look for other non-invasive diagnostic methods. Although there exists few non-invasive methods for the diagnosis of CAD like stress testing, magnetic resonance imaging (MRI) but the results are not convincing when compared to angiography[6-10]. Machine learning played a crucial role and showed its impact in almost every field of life. Medical field is also getting benefitted from it because of its capabilities to model complex relations.

Many researchers have explored the capabilities of machine learning in the past and produced many interesting and promising results. Zeinab Arabasadi et al (2017) implemented genetic algorithm for determining initial weights and neural network to present a framework for the detection of CAD. Author worked with Zalizadeh Sani data set with 303 instances to compute information gain, Principal Component Analysis. The proposed method proved efficient in terms of accuracy [11].

Babis et al (2017) worked on three different data sets: South Afri-can Heart Disease, Heart Disease Database and dataset from Z-Alizadeh Sani. They performed predictive analysis based on SVM, Decision Trees, Neural Networks and Naive Bayes. Further, they extended their research of predictive analysis on association and decision rules. The models proposed by authors using this study are comparable with existing studies and in some cases comparable or better [12].

Revised Manuscript Received on December 22, 2018.

Varun Sapra, Ph.D Scholar, Department of Computer Science, Jagannath University, Jaipur, India.

Dr. Madan Lal Saini, Department of Computer Science, Jagannath University, Jaipur, India.

Verma et al. (2016) proposed another method for coronary disease diagnostic using non-invasive clinical parameters. They proposed a blended approach where author reduced the feature space using correlation feature subset (CFS) for reducing dimensionality of the biomarkers with particle swarm optimization (PSO) search. Further, they implemented classification techniques and their method improved the prediction accuracy of diagnostic models [13].

Lin et al. (2015) in their study presented a method for reducing feature space using a Particle Swarm Optimization (PSO) which is one of the nature aspiring algorithms and Artificial Bee Colony algorithm (ABC). Their study enhanced the prediction accuracy and seemed a promising method to reduce the feature set and the identification of significant features [14].

Chen et al. worked on the prediction of financial markets. For the purpose, they used time series data of Taiwan Stock Index Futures. They implemented different feature extraction methods, and various technical indicators and expert rules for the extraction of features. They proposed planar feature demonstration methods and deep convolutional neural networks (CNN) for the improvement of trading framework. Their study proved that the use of deep learning models for financial data is an effective way and may have greater prospects [15].

P. Melillo et al. (2015) illustrated the importance and relevance of various DM techniques like artificial neural networks (ANN), Support Vector Machine (SVM) and Random Forest for increasing the prediction accuracy of cardiovascular diseases. Machine learning classifiers has shown promising outputs as compared to echo graphic parameters. Their study proved that among all the stated classifiers, performance of Random Forest is better [16].

Domínguez et al. (2017) discussed the importance of appropriate tuning of hyper-parameters for successful classification. The authors used many methods like random search, bio-inspired meta-heuristics and estimation of distribution algorithms. Their objective was to find optimal method among those who have reported good results. They worked with Firefly algorithm, Boltzmann-UMDA and Univariate Marginal Distribution Algorithm (UMDA). The study that was carried out on fifteen different medical problems proved that estimation of distribution algorithms proved as the best optimal strategy [17].

II. DATA DESCRIPTION

Z-Alizadesh Sani Dataset is used to construct the Coronary artery disease detection models. The dataset consists of 54 attributes and record of 303 subjects.

The feature recorded are organized into four categories 1) Demographic features such as age, gender, height, Body mass index, hypertension, Congestive heart failure, Airway disease, Obesity, current smoker, Ex-smoker, chronic renal failure, Cerebrovascular Accident and Dyslipidemia. 2) Echo and Laboratory features such as Lymphocyte, Platelet, Fasting blood sugar, Triglyceride, hemoglobin, Low density lipoprotein, Blood urea nitrogen, potassium, erythrocyte sedimentation rate, Sodium, White blood cell, Neutrophil, Creatine, Ejection fraction

, region with regional wall motion abnormality, valvular heart disease. 3) ECG Features such as Rhythm, Poor R progression, Left ventricular hypertrophy, T inversion, ST Depression, ST Elevation, Q wave. 4) Symptom and Examination features such as Low threshold angina, Exertional chest pain, nonanginal chest pain, Atypical, Function class, Dyspnea, Typical chest pain, Diastolic murmur, Systolic murmur, Lung rales, Weak peripheral pulse, Edema, pulse rate, blood pressure.

III. METHODOLOGY

Preprocessing is a crucial step as data recorded form heterogeneous sources can contain missing values, ambiguity and incompleteness. The prominent reason can be data transmission error, errors while doing manual data entry and faulty data collection tools. Preprocessing of the data was carried out using feature subset selection followed by model construction such as Random forest, Logistic regression, Fuzzy unordered rule induction algorithm and Multi layer perception. The performance measures that were recorded, to analyze the proposed models were accuracy, misclassification error rate, Mean absolute error and Mean Squared error.

In order to speed up the learning process dimensionality of the feature space is reduced by using correlation based feature subset selection method with firefly search. There are approximately two thousand species of firefly. Most of them produce short, inimitable and rhythmic flashes, which are a prominent source for attracting the prey or mating partners. Both male and female responds to different signals that are generated through these flashes [18]. These signals observe inverse square law i.e. the light intensity decreases with the increase in distance [19].

Feature selection technique improves the execution speed of a learning scheme; enhance data quality, performance and exploration of results. The best feature subset selects the least number of features that are highly significant and contribute maximum in terms of accuracy and efficiency of algorithm. Correlation based feature subset (CBFS) method evaluate the merit of feature subset with correlation based heuristic. The aim of CBFS is to identify a subset that is highly correlated with the outcome class label. Out of fifty four features following fifteen features have been selected for the construction of model.

Learning Schemes

Random forest is an ensemble method that can be used to solve both classification and regression problems. The classifier contains multiple tree structured models, k_1,k_2….,k_m with the goal of produce an enhanced compound classification model, k. S is the data set under consideration and is used to create m training sets, S_1,S_2,…..,S_m where S_i ($1 \leq i \leq m-1$) is used to produce classifier S_i.

Table 1: Description of Reduced Dataset

S.no	Feature	Min-max	Mean	St.Dev
1	Age	30-86	58.89	10.392
2	Diabetes Mellitus	0-1	0.297	0.458
3	Hyper tension	0-1	0.591	0.493
4	Chronic renal failute	N,Y	-	-
5	Blood pressure	90-190	129.554	18.938
6	Chest Pain (Typical)	N,Y	-	-
7	Chest pain (Atypical)	N,Y	-	-
8	Nonanginal Chest pain	N,Y	-	-
9	Q-Wave	0-1	0.053	0.224
10	Tinversion	0-1	0.297	0.458
11	Fasting blood sugar	62-400	119.185	52.08
12	Erythocyte sedimentation rate	1-90	19.462	15.936
13	Potassium	3-6.6	4.231	0.458
14	Ejection fraction	15-60	47.231	8.927
15	Region with RWMA	0-4	0.62	1.133
16	CAD	Yes, No	-	-

The greater the number of tree structured models, the higher will be the accuracy. When a new data is to be classified, it is done by each tree structured classifier by casting their vote for a class prediction.

The predictions of each tree depends on the random vector values that are sampled within the same distribution. The random attributes have been selected at each node for the purpose of split, while creating the decision trees. When classification is performed, each tree participates and the most popular class is returned. The generalization of error consists of two segments: the strength of the individual classifiers and the correlation between them [20].

An Artificial Neural Network (ANN) is a mathematical computing system that is inspired by the structure of biological neurons consists of interconnected elements, which are arranged, in multiple layers. The ANN models are best suited for large amounts of data, where inputs are limited. One of the biggest advantages of these networks that made them popular is their ability to learn from datasets. ANN has three layers, the outermost layer is the input layer where the number of neurons depends on the number of input variables. The middle layer is the hidden layer where actual computations are carried out and the result is passed on to the output layer.

Multilayer perceptron which now-a-days used with deep learning models, especially in numerical datasets becomes a widely used ANN architecture with back propagation [21]. Here we have used a deep learning sequential model with keras. The model has five layers and at each layer the weights have been assigned uniformly. The loss function used is binary cross entropy for calculating the loss.

Fuzzy Unordered Rule Induction is considered as an expansion for a RIPPER algorithm [22]. The method is used for generating fuzzy classifiers rather than conventional rules, so as to provide flexibility in modelling decision boundaries. The technique works with fuzzy set of intervals rather than conventional intervals. It makes use of membership function called trapezoidal function along with the sophisticated rule induction techniques that have been originally employed by the RIPPER algorithm

Multinomial Logistic Regressing (MLR) is a method that employs ridge estimator and is considered as an extension of logistic regression [23]. It is generally used for multi classification where we have multiple classes. Like binary logistic regression, it makes use of maximum likelihood estimation for the evaluation of the probability of categorical membership.

IV. PERFORMANCE MEASURES

The models were evaluated on the basis of the confusion matrix that is the outcome of the classification model. Classification accuracy, error rate, Mean squared error, Mean absolute errors are recorded.

True Positive (TP) –True positive is the measure of the total number of positive CAD cases that are predicted as positive by the diagnostic model.

True negative (TN) – False negative is the measure of the total number of negative cases (no CAD) that are predicted as negative (no CAD) by the diagnostic model.

False positive (FP) – False positive is the measure of total number of negative cases that are predicted as positive (CAD) by the diagnostic model.

False negative (FN) – False negative is the measure of total number of cases that are positive and classified as negative by the diagnostic model.

Accuracy- It is the percentage measure of correctly classified objects by the diagnostic model.

$$Accuracy = \frac{TP+TN}{m}$$

Error Rate – It is the percentage measure of wrongly classified object by the diagnostic model.

$$Error\ Rate = \frac{FP+FN}{m}$$

where m is the sum of the total number of cases under observation.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. RESULTS

The results of the various machine-learning methods applied to medical data set are demonstrated in Table. 2 and Figure 2. The result clearly shows that the performance of MLP based deep learning model has outperformed other diagnostic models. Further, the graphs of deep learning model that has outperformed among all are shown in Figure 3-5. Figure 3 illustrates the Accuracy of the model. Figure 4 shows the graph for Mean Absolute error and Figure 5 demonstrates the graph of Mean squared error.

Table 2: Accuracy/error rate/MAE/RMSE

Model	Accuracy (%)	Error(%)	MAE	RMSE
Deep learning	89.79	10.21	0.254	0.**112**
Random Forest	86.79	13.21	0.220	0.317
Logistic Regression	86.13	13.87	0.173	0.3136
FURIA	84.15	15.85	0.166	0.313

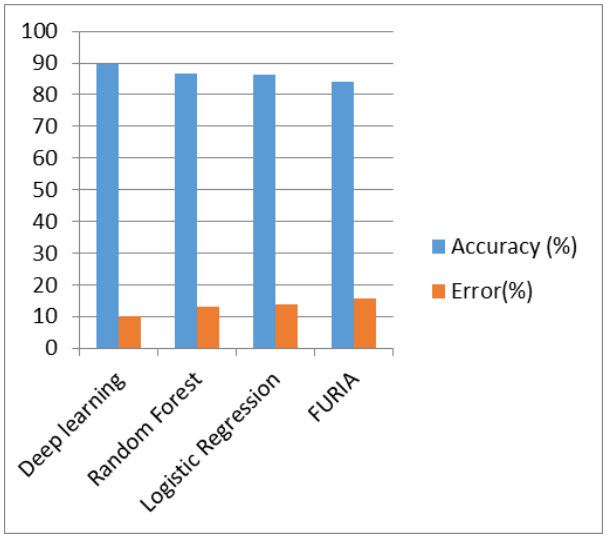

Figure 2: Accuracy and Error Rate of models using firefly search

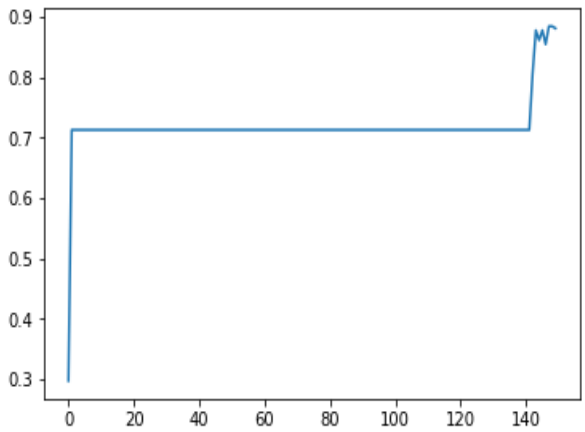

Figure 3: Accuracy of Deep learning model

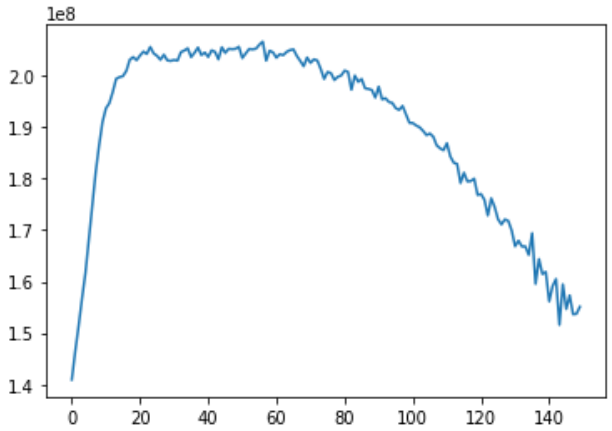

Figure 4: Mean Absolute Error for Deep learning model

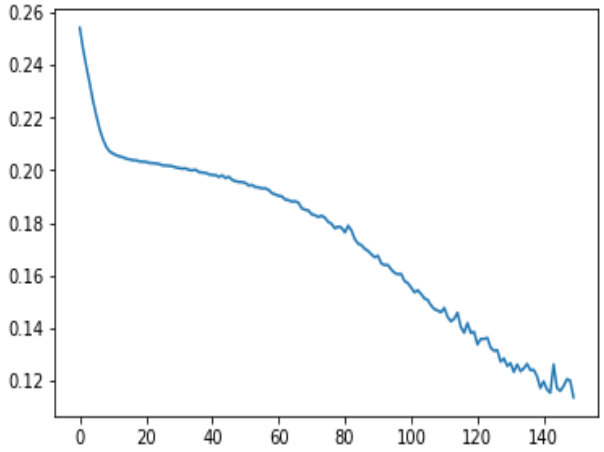

Figure 5: Mean Squared Error for Deep learning model

VI. CONCLUSION & FUTURE WORK

From the literature it has been observed that the prediction performance of an algorithm depends on different factors such as the method of data preprocessing, nature of problem, applicability of algorithm, model selection, training, testing and validation method. The work examines the use of machine learning approaches such as Fuzzy Unordered Rule Induction, Logistic regression, Random Forest and Deep learning-based classification method for Coronary Artery Disease detection. We also applied correlation-based feature subset selection method with firefly search method to reduce the dimensionality of the data set. Only fifteen parameters are used to construct the model out of fifty-four. Keras library is used to develop Multilayer perception deep learning sequential model. Prediction accuracy achieved by Random Forest is 86.79%, in case of logistic regression prediction accuracy is 86.13% that is higher than Random Forest and FURIA has the lowest prediction accuracy and highest error rate of 84.15% and 15.85%. Deep learning-based model achieved the highest prediction accuracy of 89.77% and lowest error rate of 10.21%.

Thus, deep learning-based model can be used to predict CAD cases more accurately than other schemes. Firefly search method is used to find out the most influential risk factors for identification of coronary artier disease using non-invasive clinical data. Result of deep learning based method is promising and can be used as an adjunct tool for clinical practices.

REFERENCES

1. International Heart Protection Summit, September (2011) Cardiovascular diseases in India: Challenges and way ahead. India: ASSOCHAM.
2. Randa El-Bialy, Mostafa A. Salamay, Omar H. Karam and M.Essam Khalifa, Feature Analysis of Coronary Artery Heart Disease Data Sets, International Conference on Communication, Management and Information Technology , Procedia Computer Science 65, (2015), pp:459-468.
3. Chung, J. (2017), Association between Carotid Artery Plaque Score and SYNTAX Score in Coronary Artery Disease Patients. General Medicine: Open Access, 5(5).
4. Zhou, H., Wang, X., Zhu, J., Fish, A., Kong, W., & Li, F. et al. (2017). Relation of Carotid Artery Plaque to Coronary Heart Disease and Stroke in Chinese Patients: Does Hyperglycemia Status Matter?. Experimental And Clinical Endocrinology & Diabetes, 126(03), 134-140.
5. Ceponiene, I., Nakanishi, R., Osawa, K., Kanisawa, M., Rahmani, S., & Nezarat, N. et al. (2017), Association of Coronary Artery Calcium Progression with Coronary Plaque Progression Determined by Quantitative Coronary Artery Plaque Analysis. Journal of The American College of Cardiology, 69(11), 1552. doi: 10.1016/s0735-1097(17)34941-0
6. Acharya, U., Sree, S., Muthu Rama Krishnan, M., Krishnananda, N., Ranjan, S., Umesh, P., & Suri, J. (2013). Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images. Computer Methods And Programs In Biomedicine, 112(3), 624-632.
7. Escolar E, Weigold G, Fuisz A, Weissman NJ. (2006) New imaging techniques for diagnosing coronary artery disease. Canadian Medical Association Journal. 2006 Feb 174(4), pp. 487-95.
8. Giri D, Acharya UR, Martis RJ, Sree SV, Lim TC, Ahamed T, Suri JS. Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform. Knowledge-Based Systems. 2013 Jan 31;37, pp.274-282.
9. Alizadehsani R, Habibi J, Hosseini MJ, Mashayekhi H, Boghrati R, Ghandeharioun A, Bahadorian B, Sani ZA. (2013), A data mining approach for diagnosis of coronary artery disease. Computer methods and programs in biomedicine. 2013 Jul 31;111(1), pp. 52-61
10. Kahramanli, H., & Allahverdi, N. (2008), Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications, 35(1-2), 82-89.
11. Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., & Yarifard, A. A. (2017), Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Computer methods and programs in biomedicine, 141, 19-26.
12. Babič, F., Olejár, J., Vantová, Z., & Paralič, J. (2017), Predictive and descriptive analysis for heart disease diagnosis. In Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on (pp. 155-163). IEEE.
13. Verma, L., Srivastava, S., & Negi, P. C. (2016), A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. Journal of medical systems, 40(7), 178.
14. Lin, K. C., & Hsieh, Y. H. (2015), Classification of medical datasets using SVMs with hybrid evolutionary algorithms based on endocrine-based particle swarm optimization and artificial bee colony algorithms. Journal of medical systems, 39(10), 119.
15. Jou-Fan Chen, Wei-Lun Chen, Chun-Ping Huang, Szu-Hao Huang, An-Pin Chen (2016), Financial Time-series Data Analysis using Deep Convolutional Neural Networks, 7th International Conference on Cloud Computing and Big Data, p.p 87-92.
16. P. Melillo, R. Izzo, A. Orrico, P. Scala, M. Attanasio, M. Mirra, N. De Luca and L. Pecchia (2015), Automatic Prediction of Cardiovascular and Cerebrovascular Events Using Heart Rate Variability Analysis, PLOS ONE, vol. 10, no. 3, p. e0118504, 2015.
17. Alfonso Rojas-Domínguez, Luis Carlos Padierna, Juan Martín Carpio Valadez, Hector J. Puga-Soberanes, Héctor J. Fraire (2017), Optimal Hyper-Parameter Tuning of SVM Classifiers With Application to Medical Diagnosis, IEEE Access, Vol 6, 2018, p.p. 7164-7176.
18. X.S. Yang (2010), Firefly algorithm, levy flights and global optimization in Research and Development in Intelligent Systems XXVI. Springer, pp. 209–218.
19. X.S. Yang (2011), Chaos-enhanced firefly algorithm with automatic parameter tuning, International Journal of Swarm Intelligence Research (IJSIR), Vol. 2, No. 4, pp. 1–11.
20. Han, J., & Kamber, M. (2011). Data mining: Concepts and techniques (3rd ed.). Burlington, MA: Elsevier.
21. Cohen, W. W. (1995), Fast effective rule induction. In Proceedings of the twelfth international conference on machine learning, pp. 115–123.
22. J. Hühn, and E. Hüllermeier (2009), FURIA: an algorithm for unordered fuzzy rule induction. Data Mining and Knowledge Discovery, vol. 19, no. 3, pp. 293-319.
23. Le Cessie, and J Houwelingen (1992), "Ridge Estimators in Logistic Regression", Applied Statistics, vol. 41, no. 1, pp. 191-201.