International Journal of Applied Engineering Research ISSN 0973-4562 Volume 16, Number 8 (2021) pp. 685-689
© Research India Publications. http://www.ripublication.com

A Hybrid Classification Method for Heart Disease Detection

Swapnil Ashtekar[1], **Pranit Kotkar**[1] **and Sakshat Patil**[1]

Vidyalankar Institute of Technology Mumbai, India.

Abstract

The hybrid model development technique is penetrating the newest technological developments and has emerged as part of the analysis. It's gaining vast traction for planning AI solutions. The work applied during this paper is to develop a hybrid model which may predict heart disease expeditiously. Heart disease or cardiovascular disease (CVD) is a primary reason behind death worldwide, and the traditional diagnostic ways like electrocardiography (ECG) or Electron-Beam Computed Tomography (EBCT) or coronary angiography (CA), etc. may be defective. Also, detection of those vast diseases could be a cumbersome task for a physician. To assist physicians, create fast choices and minimize errors in identification, hybrid systems change physicians to look at the medical knowledge in substantial detail. We tend to use the present model of logistic regression and collaborated with artificial neural networks within the prediction. The results proved to be outstanding after integrating both models compared to once singly enforced.

Keywords: Hybrid Classification, Heart Disease Detection, Neural Networks, Logistic Regression

INTRODUCTION

Heart disease or cardiovascular disease (CVD) could be reasonably ill health that involves the heart and/or blood vessels of individuals throughout the world [1]. Cardiovascular could be a primary explanation of death worldwide. Per the World Health Organization, India accounts for twenty percent of those deaths worldwide particularly in a younger population [2]. Therefore, accurate diagnosing of heart disease within the early stages is of great significance in raising the safety of the heart [3].

To discover heart disease, many diagnostic methods ways are developed such as Electrocardiogram (ECG), Coronary Angiography (CA), Electron-Beam Computed Tomography (EBCT), Cardiac Catheterization, Echocardiography, Cardiac MRI, etc. [4], however, these diagnostic methods can have serious defects. To date, two-thirds of all heart attacks go unobserved on Electrocardiogram [5] whereas methodology like Coronary Angiography (CA) may be invasive and needs accomplished operators [6]. Thus, detection of those cardiovascular diseases could be a cumbersome task for a physician. To assist physicians, build fast selections, and minimize errors in diagnosing, classification systems modify physicians to rapidly examine medical information in extensive detail [7]. These systems are enforced by developing a model that may classify existing records using sample data. Numerous classification algorithms are developed and used as classifiers to help doctors diagnose heart disease patients.

Detection of heart disease could be a crucial task; therefore, the classification model must have, high accuracy. Compared to classifiers operating one by one, classifiers operating along can have a higher potential for gaining better accuracy [8]. Diverse classifiers operating along can have a higher potential for gaining better accuracy compared to non-diverse classifiers operating along [9].

Some researchers use a combination of various varieties of classification algorithms to create hybrid models. As an example, concerning a hybrid model composed of SVM, KSVM, and artificial neural networks, P. Hemalatha et al. use such a model in customer church prediction [10]. A.D. Kumar et al. use various algorithms such as multilayer perceptron, RBF network, etc. to predict the academic performance of a student [11]. Bharat et al. use integrations of SVM and J48 for text classification [12]. Furthermore, Edward J. Kim et al. uses TPC, SOMC, SOMs, and HB for star-galaxy classification [13]. Here, hybrid models come through higher classification performance compared to non-hybrid models.

In this research, the emphasis lies in studying data and building a hybrid classification model to detect heart disease. It requires a deep dive into several machine learning supervised algorithms and testing them to identify the best model aligning with the research objective.

DATASET

Data Collection

For this research, the Heart Disease UCI dataset was chosen, and therefore the numerous fields of the dataset are that of patient profiles. The Cleveland (Cleveland Clinic Foundation) database is chosen for this research as a result of its usually used database for machine learning researchers with comprehensive and complete records.

Data Exploration

The dataset includes 300+ samples with a total of 14 medical features, their description are given in Table 1. The feature

"Class" contains whether or not a patient incorporates a presence or absence of heart disease.

Table 1. Dataset Attributes Description

Sr. No	Code	Feature	Description
1.	Age	Age	Age in years
2.	Sex	Sex	sex (1 = male; 0 = female)
3.	Cp	Chest pain type	1 = typical angina; 2 = atypical angina; 3 = non-angina pain; 4 = asymptomatic
4.	Trestbps	Resting blood pressure (mg)	At the time of admission in hospital [94, 200]
5.	Chol	Serum cholestrol (mg/dl)	Multiple values between [Minimum Chol: 126, Maximum Chol: 564]
6.	Fbs	Fasting blood sugar > 120 mg/dl	1 = yes; 0 = no
7.	Restecg	Resting electrocardiographic results	0 = normal; 1 = ST-T wave abnormal; 2 = left ventricular hypertrophy
8.	Thalach	Maximum heart rate achieved	Maximum heart rate achieved [71, 202]
9.	Exang	Exercise-induced angina	1 = yes; 0 = no
10	Oldpeak	ST depression induced by exercise relative to rest	Multiple real number values between 0 and 6.2
11.	Slope	The slope of the peak exercise ST segment	1 = unsloping; 2 = flat; 3 = downsloping
12.	Ca	Number of major vessels (0-3) colored by fluoroscopy	Number of major vessels colored by fluoroscopy (values 0-3)
13.	Thal	Exercise thallium scintigraphy	3 =normal; 6 = fixed defect; 7 = reversible defect
14.	Class (Target)	The predicted attribute	0 = no presence; 1 = presence

Data Preprocessing

In the dataset, sex, cp, fbs, restecg, exang, slope, ca, thal are categorical columns. Therefore, for the application of machine learning algorithms, the columns had to be encoded. MinMaxScaler, imported from the sklearn library, was used for feature scaling of the input data.

PROPOSED WORK

Early and accurate detection and diagnosis of cardiac diseases are of extreme importance. Reliable conventional techniques for heart detection include invasive coronary angiography or cardiac catheterization, non-invasive methods such as Electrocardiogram (ECG), Cardiac Computerized Tomography (CT), and Cardiac Magnetic Resonance Imaging (MRI) can sometimes be deceitful. This work proposes an intelligent learning-based computational method to assist in diagnosing heart diseases or give an early warning of the presence of the same. Fig. 1. demonstrates the proposed framework in a data flow representation. In the proposed work, created the training data and the validation data using the pre-processed imported data. Classification using Logistic Regression then validate and predict the data set, assigning the predicted values achieved in Logistic Regression classification to the target values for classifying using ANN. Then finally, validation and prediction to detect the presence of heart disease.

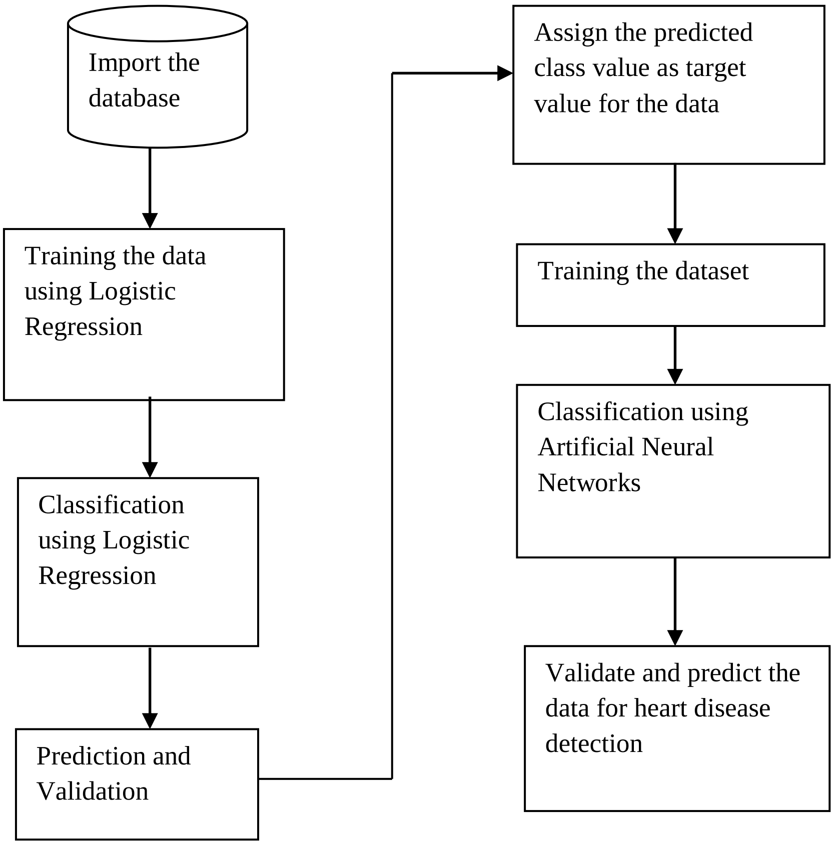

Import the database
Training the data using Logistic Regression
Classification using Logistic Regression
Prediction and Validation
Assign the predicted class value as target value for the data
Training the dataset
Classification using Artificial Neural Networks
Validate and predict the data for heart disease detection

Fig.1. Proposed Work

EXPERIMENTS AND RESULTS

This work is implemented using Jupyter Notebook version 6.1.4.

Classification using Logistic Regression Algorithm

Utilized Scikit-learn's Logistic Regression API to classify the data. Logistic regression is a statistical model that uses a logistic function to model a binary dependent variable. It calculates the odds of the event for different levels of each independent variable, then takes its logarithm to create a continuous criterion as a transformed version of the dependent variable. An accuracy of 84.78% has been obtained for the logistic regression model while using the K-fold Cross-Validation technique.

Classification using ANN Algorithm

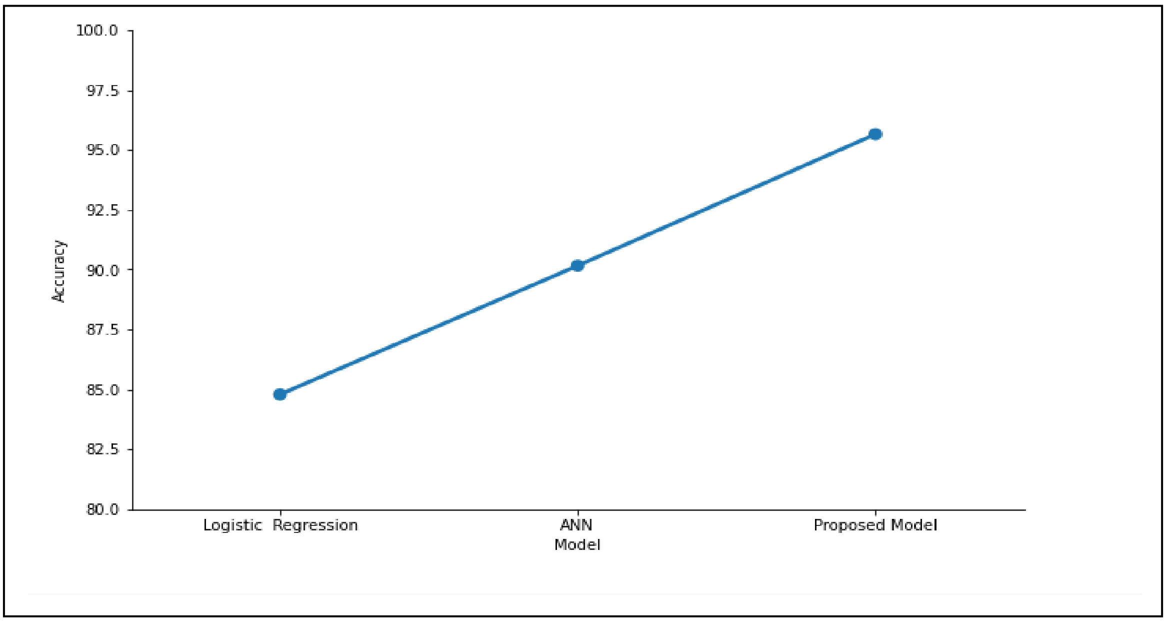
Used Keras sequential API to create a Feed-Forward Neural Network model with two hidden layers and one output layer. The twelve nodes' first hidden layer used the ReLU activation function, the twelve nodes' second hidden layer used the ReLU activation function, and the output layer had one node and used the Sigmoid activation function. The training accuracy received after the first run of the model was 90.16 %, but re-runs of the model caused the accuracy to be inconsistent. The performance of the model using only ANN could have been increased and made consistent with more data.

Classification using Logistic Regression and ANN Algorithm

The adopted hybrid model of Logistic Regression and ANN showed better accuracy results and, it even improved the model when examined the metrics with that of traditional classifiers. The hybrid combination resulted by fetching the output produced by the Logistic Regression and running the model with ANN classifier resulting in better model prediction values of heart data in the database acquired.

Table 2. Model Evaluation

Models	**Logistic Regression**	**ANN**	**Logistic Regression + ANN**	**Derivations**
(Measure)		***(Value)***		
Sensitivity	0.9117	0.8235	0.9600	TPR = TP/(TP+FN)
Specificity	0.8518	0.8888	0.9047	SPC=TN/(FP+TN)
Precision	0.8857	0.9032	0.9230	PPV=TP/(TP+FP)
Negative Predictive Value	0.8846	0.8000	0.9500	NPV=TN/(TN+FN)
False Positive Rate	0.1481	0.1111	0.0952	FPR=FP/(FP+TN)
False Discovery Rate	0.1142	0.0967	0.0769	FDR=FP/(FP+TP)
False Negative Rate	0.0882	0.1764	0.0400	FNR=FN/(FN+TP)
Accuracy	0.8478	0.9016	0.9565	ACC=(TP+TN)/(P+N)
F1 Score	0.8985	0.8615	0.9411	F1=2TP/(2TP+FP+FN)
Matthews Correlation Coefficient	0.7932	0.7467	0.8765	TP*TN – FP*FN / sqrt((TP+FP) *(TP+FN) *(TN*FP) *(TN+FN))



CONCLUSION

Heart disease or cardiovascular disease (CVD) is a primary cause of death worldwide, and the traditional diagnostic methods such as electrocardiography (ECG) or Electron-Beam Computed Tomography (EBCT) or coronary angiography (CA), etc. can be defective. Also, detection of these cardiovascular diseases is a cumbersome task for a physician. To help physicians make quick decisions and minimize errors in diagnosing, we proposed a hybrid system that will enable physicians to examine the medical data in considerable detail. A broad arrangement of regression affirms that the proposed approach displays 92.30 % precision compared to the implementation of ANN or Logistic Regression individually. As seen in fig. 2. the accuracy achieved by the hybrid model is the highest. Therefore, successfully implemented the hybrid model composed of ANN and Logistic Regression to predict heart disease from the database acquired.

REFERENCES

[1] Biglu, M. H., Ghavami, M., & Biglu, S. (2016). Cardiovascular diseases in the mirror of science. *Journal of cardiovascular and thoracic research*, *8*(4), 158–163. https://doi.org/10.15171/jcvtr.2016.32

[2] Sreeniwas Kumar, AT., & Sinha, N. (2020). Cardiovascular disease in India: A 360 degree overview. *Medical journal, Armed Forces India*, *76*(1), 1–3. https://doi.org/10.1016/j.mjafi.2019.12.005

[3] Amato F, Lopez A, Pena-Mendez EM, Vanhara P, Hampl A, Havel J. Artificial neural networks in medical diagnosis. J Appl Biomed. 2013;11(2):47–58.

[4] *Cardiovascular Disease Diagnosis.* [online] news-medical.net. Available at: < https://www.news-medical.net/health/Cardiovascular-Disease-Diagnosis.aspx> [Accessed 14 September 2021]

[5] *ECG often fails to diagnose heart attacks.* [online] enverdis.com. Available at: < https://www.enverdis.com/blog/2011/07/01/ecg-often-fails-to-diagnose-heart-attacks/> [Accessed 14 September 2021]

[6] Tavakol, M., Ashraf, S., & Brener, S. J. (2012). Risks and complications of coronary angiography: a comprehensive review. *Global journal of health science*, *4*(1), 65–93. https://doi.org/10.5539/gjhs.v4n1p65

[7] H. Temurtas, N. Yumusak, F. Temurtas, A comparative study on diabetes disease diagnosis using neural networks, Expert Systems with Applications, vol. 36, no. 4, pp. 8610-8615, 2009, https://doi.org/10.1016/j.eswa.2008.10.032.

[8] H. T. X. Doan & G. M. Foody (2007) Increasing soft classification accuracy through the use of an ensemble of classifiers, International Journal of Remote Sensing, 28:20, 4609-4623, DOI: 10.1080/01431160701244872

[9] L. I. Kuncheva, M. Skurichina, and R. P. W. Duin, “An experimental study on diversity for bagging and boosting with linear classifiers,” *Information Fusion*, vol. 3, no. 4, pp. 245–258, 2002, https://doi.org/10.1016/S1566-2535(02)00093-3.

[10] P. Hemalatha and G. M. Amalanathan, "A Hybrid Classification Approach for Customer Churn Prediction using Supervised Learning Methods: Banking Sector," *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, 2019, pp. 1-6, doi: 10.1109/ViTECoN.2019.8899692.

[11] A. D. Kumar, R. P. Selvam and V. Palanisamy, "Hybrid Classification Algorithms for Predicting Student Performance," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2021, pp. 1074-1079, doi: 10.1109/ICAIS50930.2021.9395974.

[12] Singh, Bharat & Kushwaha, Nidhi & Vyas, O.. (2016). A Scalable Hybrid Ensemble Model for Text Classification. 10.1109/TENCON.2016.7848630.

[13] Edward J. Kim, Robert J. Brunner, Matias Carrasco Kind, A hybrid ensemble learning approach to star–galaxy classification, *Monthly Notices of the Royal Astronomical Society*, Volume 453, Issue 1, 11 October 2015, Pages 507–521, https://doi.org/10.1093/mnras/stv1608.