Development of a Deep Learning-Based Predictive Model for Early Detection and Risk Assessment of Cardiovascular Diseases in Patients with Diabetes Mellitus

Alamma B.H.
Assistant Professor, Dept. of MCA, Dayananda Sagar College of Engineering,
Bangalore-560072, Karnataka, India
&
VTU Ph.D. Research Scholar (Part-Time), Dept. of MCA,
Sir M. Visvesvaraya Institute of Technology, Bangalore, Karnataka
Email : alamma-mcavtu@dayanandasagar.edu

Dr. Manjula Sanjay Koti
Supervisor & Professor – Head of the Dept., Dept. of Master of Computer Applications (MCA),
Dayananda Sagar Academy of Technology and Management, Bangalore-560082, Karnataka
Email : manjula.dsce@gmail.com

Dr. C.H. Vanipriya
Co-Supervisor, Professor & HOD, MCA Dept., Sir M. Visvesvaraya Institute of Technology, Krishnadevaraya Nagar, Hunasamaranahalli, International Airport Road, Bangalore - 562157
Email : vanipriya.manmohan@gmail.com

Abstract - In recent years, medical technology has grown by leaps and bounds, resulting in the generation of an enormous amount of patient-related data, especially during clinical trials. These trials often produce information that spans multiple dimensions—covering everything from lab results to patient histories. This explosion of data has become the foundation for big data analytics in healthcare, enabling researchers and clinicians to uncover hidden patterns and risk factors behind serious conditions like cardiovascular and respiratory diseases. With the rise of intelligent systems, robots equipped with advanced analytics are increasingly supporting doctors in detecting illnesses early and more accurately, especially in complex scenarios where human observation alone might miss subtle signs. Among the various life-threatening conditions, cardiovascular disease remains one of the leading causes of death globally. Given its widespread nature, many data-driven approaches have been explored to predict and manage its risks. In this study, we focus on how diabetes—a known contributor to cardiovascular complications—interacts with these diseases. We employ a deep learning neural network model to analyze patient data and reveal these connections. Our findings show that the model outperforms many existing methods in terms of accuracy and F1 score, demonstrating its potential to enhance early diagnosis and provide actionable insights for preventive healthcare strategies.

Keywords: Data Analytics, Cardio Vascular Diseases, Diabetes, HealthCare.

1. Introduction

The rapid growth of information and communication technology has had a profound impact on the healthcare industry, making diagnosis and monitoring more accessible and affordable for all types of patients. By providing timely updates and health insights, it has empowered individuals to take a more proactive role in managing their well-being. According to the World Health Organization (WHO), cardiovascular diseases accounted for 32% of global deaths in 2019. In India alone, the Ministry of Health and Family Welfare reported that 28.1% of all deaths were due to cardiovascular issues. These alarming figures highlight the urgency of addressing heart-related illnesses, which continue to be the leading cause of death worldwide.

With the growing availability of medical data, machine learning has emerged as a powerful tool in predicting health outcomes with greater accuracy. Its ability to classify diseases more effectively has led to widespread adoption across sectors such as healthcare, transportation, social media, and more. In the medical field, machine learning and data mining techniques are now routinely used to analyze extensive patient records, providing valuable insights that were previously difficult to obtain. Factors like a high-fat diet, elevated blood pressure, poor cholesterol levels, sedentary lifestyles, and genetic predispositions are all contributors to cardiovascular disease. By analyzing these variables, researchers can develop models that forecast the risk of heart disease, potentially transforming how the medical community approaches prevention and treatment, ultimately improving patient outcomes and saving lives [10].

2. Literature Survey

The authors classified heart illness using Logistic Regression (LR) methods using data from UCI. To enhance the performance of the model, the dataset was cleaned, missing values were located, and correlation with the target value was performed for all characteristics in order to choose those that were highly positive correlated. The dataset is split into training and testing groups in the following ratios: 90:10, 80:20, 70:30, 40:60, and 50:50. The highest accuracy 87.10%.was achieved with a splitting ratio of 90:10 [1]. Extra Trees Classifier, Random Forest, XGBoost, and other machine learning techniques are used in a framework developed by the authors of this paper that has a stacked ensemble classifier. Several performance indicators were applied to the proposed model in order to evaluate its effectiveness and robustness. They have surpassed the available literature with an accuracy of 92.34% [2].

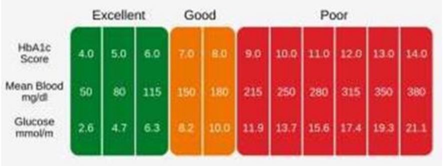
Based on one of the key features, such age, the authors of this study have shown how to forecast and analyse heart-related syndromes in patients. By doing this, data scientists can use big data to conduct early analyses of heart syndromes and perhaps save patients' lives. In this case study, numerous well-considered variables are employed to analyse and predict heart illnesses in patients. The author then used predictions from the data to examine the accuracy of the syndrome diagnosis. The suggested system will be useful for studying cardiovascular disease [5]. In this paper, the authors discuss the prediction of CHD risk using machine learning methods such as Random Forest, Decision Trees, and K-Nearest Neighbors. Additionally, these algorithms are compared based on how accurately they make predictions. K-fold Cross Validation is also used to give the data some volatility. These techniques are examined using the 4240-record “Framingham Heart Study” dataset. In our experimental research, the accuracies of Random Forest, Decision Tree, and K-Nearest Neighbor were 96.8%, 92.7%, and 92.89%, respectively. As a result, Random Forest classification provides more accurate results than other machine learning algorithms when our preprocessing methods are used [6].

3. Methodology

The cardio vascular disease data was obtained from Spectrum Diagnostic Centre & Health Care, Bangalore. The amount of data collected was 1280 instances with 14 features. One of the parameters was devoted as class label and the remaining were taken as the primary risk factors of the Cardio vascular diseases, like diabetics and BMI. Two variables from the dataset—sex and age, are used to uniquely identify each patient record and assign individual identifiers. Medical data makes up the remaining features. The medical data is essential for identifying risk factors for heart disease.

Data Pre-processing: It is a major stage to gain more significant precision. Data preprocessing was done to cleanse the data and remove the missing data. The missing values in data set was managed using “Median” approach.

Table -1 Classification of Diabetic patience

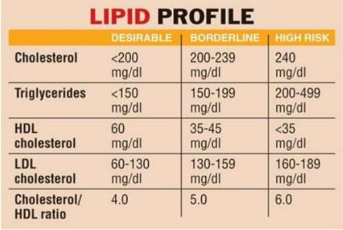


	Excellent			Good		Poor					
HbA1c Score	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0	14.0
Mean Blood mg/dl	50	80	115	150	180	215	250	280	315	350	380
Glucose mmol/m	2.6	4.7	6.3	8.2	10.0	11.9	13.7	15.6	17.4	19.3	21.1

Total	Bad	Good	VG
2128	880	392	856

From the table -1 according to the medical standard 880 records were classified as patients suffering from diabetic by comparing their HBA1C values, 392 patients were pre-diabetic and 856 patients were normal patients.

Table-2 CVD patient categorization



LIPID PROFILE

	DESIRABLE	BORDERLINE	HIGH RISK
Cholesterol	<200 mg/dl	200-239 mg/dl	240 mg/dl
Triglycerides	<150 mg/dl	150-199 mg/dl	200-499 mg/dl
HDL cholesterol	60 mg/dl	35-45 mg/dl	<35 mg/dl
LDL cholesterol	60-130 mg/dl	130-159 mg/dl	160-189 mg/dl
Cholesterol/ HDL ratio	4.0	5.0	6.0

CVD			
Total	0	1	% of CVD
2128	440	1688	79%

From table-2 Using the medical standards of lipid profile test and comparing it with the parameters LDL and HDL values 1688 patients were found suffering from CVD and 440 patients were not having CVD.

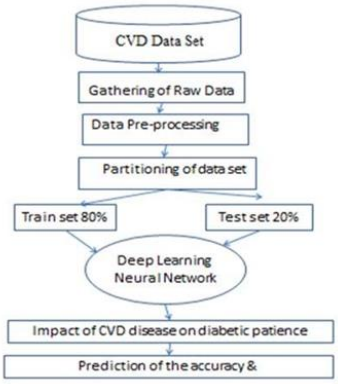


Fig.1 : Block diagram of the proposed methodology

4. Data Modelling Procedure

The partition of the data set was done as train (70%) and test (30%) sets, coupled with the computed dependent variable (CVD) Deep learning neural network was used to model these data sets. Based on performance criteria like accuracy and F1 score, this model is evaluated. The proposed methodology's block diagram is shown in Figure 1. This methodology involves gathering a CVD data set from Spectrum Diagnostic Center & Health Care and cleaning the data so that it will appropriately fit the model. After that, the data is divided into training and test sets for the deep learning neural network.

5. Deep Neural Network

Deep neural networks (DNNs) are artificial neural networks (ANNs) that have more depth, or hidden layers, between the input and output layers. Deep learning is a more well-known machine learning method. It employs standard tabular data in addition to picture classification tasks. An intriguing problem in deep learning is to learn the nonlinear mapping between the inputs and outputs as well as the underlying structure of the data (input) vectors. The process of improving a neural network’s accuracy is called training. The output of a forward prop net is compared to the known accurate value. The cost function, also known as the loss function, is the difference between the output that was generated and the output that was actually produced. The cost function or loss function is the difference between the output that was generated and the output that was actually produced. Each set of inputs can be alter’d adjusting the weights and biases assigned to each edge and node. The accuracy of a neural network’s predictions depends on its weights and biases [16]. The process of improving a neural network’s accuracy is called training. A deep neural network has more hidden layers than a traditional model.

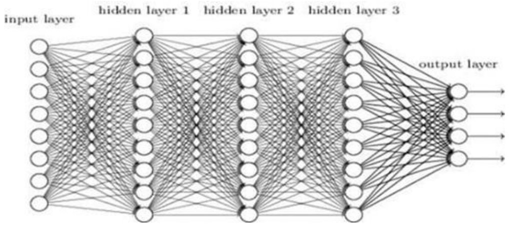


Fig. 2 : Architecture of Deep Neural network

From Figure-2, it is shown layering of nodes in a deep neural network is determined by the architecture of the network. The neural network's functional activity is largely governed by its architecture, which varies based on the application.

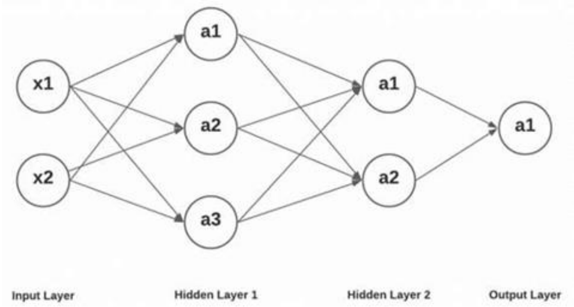


Fig. 3 : Block diagram of Deep Neural Network

From Figure-3. it is shown that the input passes through a number of hidden levels that are more than two layers as it moves from the input layer to the output layer. Three layers—the input layer, the hidden layer, and the output layer—make up the suggested model. There are a set number of neurons in each layer. This study uses a feed-forward multilayer perceptron with five input layers, one or more hidden layers, and three output layers as its neural network architecture. A vector of different features taken from the cardiovascular dataset are the input to the network. Each hidden layer gets an input vector from the layer below and transforms it using a linear transformation and nonlinear activation to create its output vector. For neuron j in hidden layer l, its output is

$$f(b + \sum^{n} x_i \ w_i)$$

where
b = bias
x = input to neuron
w = weights
n = the number of inputs from the incoming layer
i = a counter from 1 to n

The sigmoid function, or f, is the activation function. Three output layers, six input layers, and 100 epochs were used in this study.

Results and Discussions

Performance for categorization issues at various threshold levels is measured by the AUC-ROC curve. AUC stands for the level or measurement of separability, and ROC is a probability curve. It illustrates how well the model is in differentiating between classes. The model performs better at classifying 0 classes as 0, and 1 classes as 1, the higher the AUC. As a result, the model does a better job at differentiating between those who have the condition and those who do not the higher the AUC. Instructional ROC. The Training data set's AUROC is 90.54%.

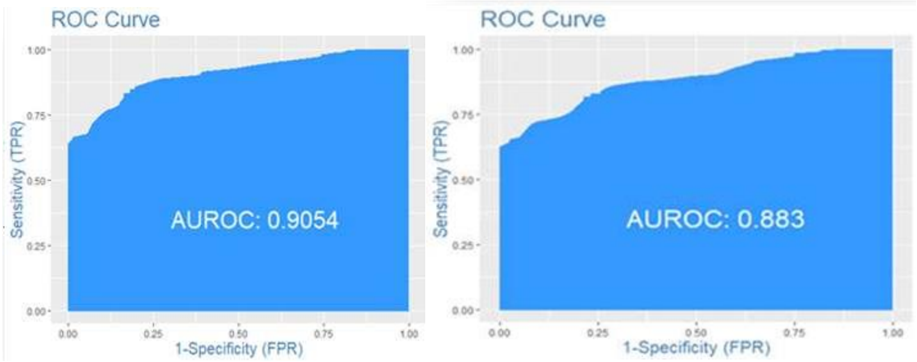


Fig. 4 : ROC curve - 1

The F1 score is a prominent performance measure for classification and is frequently preferred over, for example, accuracy when data is unbalanced, for as when the number of samples belonging to one class greatly outnumbers those found in the other class. For the Test dataset, the model's F1 score is 89%. A high level of separability, which is a hallmark of a good model, is indicated by an AUC that is close to 1. A mediocre model is indicated by an AUC that is close to 0, the worst indicator of separability. In fact, it implies that the result is being changed. It is expected that all 1s and 0s will be 1. If AUC is 0.5, the model also has no capacity for class separation. When AUC is 0.883, there is an 88.3% chance that the model will be able to distinguish between positive class and negative class. AUROC for the test data set is 88.30%.

6. Conclusions

In India, limited resources and population expansion coexist. Better healthcare is desperately needed, as demonstrated by COVID-19. Predicting cardiovascular disease using the patient's current data is one of the key facets of the medical field. Cardiovascular disease can be predicted using a variety of methods and technologies. This study uses a deep neural network to classify heart disease. To improve performance, pre-processing tasks like cleaning and locating missing values are completed. The key element is feature selection, which enhances algorithm accuracy and even concentrates on the behavior of the algorithm. In this scary scenario, the approach suggested can help with early patient diagnosis and early cardiovascular disease forecasting for the healthcare sector. Following an evaluation of current approaches, the most precise and successful strategies for predicting cardiovascular disease were identified. Given the severity of cardiovascular disease, one of the most popular areas of study nowadays is its analysis. This study showed notable improvements in cardiovascular disease detection performance. Because the accuracy and F1 score of the proposed approach are obviously higher than those of the bulk of the existing methodologies, this research has a significant influence. For the train data set, the accuracy, F1score, and AUCROC were 83.85%, 90%, and 90%; for the test dataset, the corresponding values were 82.86%, 89%, and 88.3%.

References

[1] A.G, B. Ganesh, A. Ganesh et al., "Logistic regression technique for prediction of cardiovascular disease". Global Transitions Proceedings 3 127–130. 2022

[2] Tiwari et al., "Ensemble framework for cardiovascular disease prediction ".Computers in Biology and

Medicine 146 105624", 2022.

[3] S. Dewangan.et.al. "Diabetes Prediction Using Machine Learning Techniques," Int. Journal of Engineering Research and Application. January 2018, pp.-09-13

[4] J. Azmi et al. "A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data", Medical Engineering and Physics 105 103825 2022.

[5] M.S. Nawaz et al., "Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization", *https://doi.org/10.1016/j.heliyon.2021.e06948*

[6] Noor Basha Ashok Kumar P S,et al, "Early Detection of Heart Syndrome Using Machine Learning Technique", *IEEE* 978-1-7281-3261-7/19/$31.00 © IEEE 2019

[7] Divya Krishnani, Anjali Kumari .et.al ., "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms ", 978-1-7281-1895-6/19/$31.00 c IEEE 2019

[8] Fabio Mendoza Palechor*, Alexis De la Hoz Manotas et al., "Cardiovascular Disease Analysis Using Supervised and Unsupervised Data Mining Techniques", Volume 12, Number 2, February 2017.

[9] D.Karthick, B.Priyadharshini, "Predicting the chances of occurrence of Cardio Vascular Disease (CVD) in people using Classification Techniques within fifty years of age", 978-1-5386-0807-4/18/$31.00 © IEEE, 2018

[10] Ms Fatima Dilawar MullaDr. NaveenKumar Jayakumar,., "A Review of Data Mining & Machine Learning approaches for identifying Risk Factor contributing to likelihood of Cardiovascular Diseases" 978-1-5386-4985-5/18/$31.00 © IEEE, 2018.

[11] Abderrahmane Ed-daoudy, Khalil Maalmi., "Real-time machine learning for early detection of heart disease using big data approach", 978-1-5386-7850-3/19/$31.00 © IEEE, 2019

[12] Nitten S. Rajliwall | Rachel Davey,*et.al.*, "Machine learning based models for Cardiovascular risk prediction", 978-1-7281-0404-1/19/$31.00 © IEEE, 2019.

[13] P. Suresh Kumar, S. Pranavi", "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics",*International Conference on Infocom Technologies and Unmanned Systems* (ICTUS'2017),Dec. 18-20, 2017, ADET.

[14] Sajida Perveen, Muhammad Shahbaz, Karim Keshavjee, "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on MachineLearning Techniques", pp. 2169-3536, IEEE, 2018.

[15] Sundus Abrar, Chu Kiong Loo *et.al.*, "A Multi-Agent Approach for Personalized Hypertension Risk Prediction",*Digital Object Identifier* 10.1109/ACCESS.2021.3074791

[16] Dinu A.J., Ganesan R., Felix Joseph and Balaji V,"A study on Deep Machine Learning Algorithms for diagnosis of diseases", *International Journal of Applied Engineering Research*, ISSN 0973-4562, Volume 12, Number 17. 2017.

[17] Ajad Patel, Sonali Gandhi, Swetha Shetty,Prof. BhanuTekwani, "Heart Disease Prediction Using Data Mining", *IRJET*, Vol. 4, Issue1, Jan -2017.