Implementation of deep learning algorithm to analyze the effect of diabetes on cardiovascular diseases

[1]Alamma B.H. & [2] Dr. Manjula Sanjay Koti
[1]Assistant Professor, Dept. of MCA, Dayananda Sagar College of Engineering, Bangalore-72, Karnataka, India
&
VTU Ph.D. Research Scholar (Part-Time), Dept. of MCA, Sir MVIT, Bangalore, Karnataka
[2] Superviosr, & Professor - Head, Dept. of Master of Computer Applications, Dayananda Sagar Academy of Technology and Management, Bangalore-560082, Karnataka, India
Emails : alammat@yahoo.co.in manjula.dsce@gmail.com

Abstract

Massive advancements in healthcare technology have helped to generate a tons of data about the patient's health. Data with several dimensions are produced as a result of the patients' clinical trails. The foundation for big data analytics in the healthcare sector is knowledge creation from extremely large and diverse data sets, enabling the detection of disease risk factors like cardiovascular or respiratory disorders. Instead of doctors acting as the only key aspect in determining the risk factors and diagnose of the diseases, intelligent machines can be more successful. Cardiovascular disease is one of the main diseases which is a threat to most people. Hence, data analysis and prediction model on cardiovascular disease (CVD) have been built because of its prominence. In this study, the impact of diabetes on cardio vascular disorders is examined using a deep learning neural network model. The accuracy of the model and F1 score of the suggested methodology are clearly higher than those of the majority of the current methodologies used in the literature.

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

I. Introduction

All categories of patients have benefited from the rapid development of information and communication technology, which has reduced the cost of diagnosing and monitoring patients' health by giving patients timely information [1]. The World Health Organization (WHO) reported that cardiovascular illnesses caused 32% of all deaths in 2019. Out of which 28.1% of all deaths in India were attributed to cardiovascular illnesses, according to the Ministry of Health and Family Welfare. Heart disorders are the leading global cause of death [11]. Machine learning seeks to produce increasingly favourable outcomes for precise forecasts. Therefore, it appears that machine learning is quite good in better categorising diseases. Numerous industries, including healthcare, social media, retail, and traffic control, self-driving cars, speech recognition, picture recognition, and medical diagnostics, have used machine learning applications. According to, medical data analysis and information extraction heavily utilise machine learning and data mining techniques. Due to this, patient health records are now substantially more detailed. The research has access to the extensive medical records. The use of the massive medical data presents the medical business with great challenges. Machines quickly turn the enormous volume of data to provide accurate and useful information. Machine learning is therefore a crucial field. The biggest challenge facing the medical sector today is to upgrade health infrastructure with better facilities so that diseases can be detected early on and treated promptly to enhance quality of life. One of the most serious diseases in the world is cardiovascular disease (CVD). According to medical studies, a person's chance of developing cardiovascular disease, especially heart disease, is influenced by a number of health conditions.

These include a high-fat diet, high blood pressure, low HDL (good) cholesterol, high LDL (bad) cholesterol, a history of CVDs in the family, and a lack of regular exercise. The medical community and people's lives may both be significantly impacted by the ability to forecast cardiovascular disease [10]. Numerous elements contribute significantly to the problem of cardiovascular disease (CVD). Misdiagnosis is one of the main causes of the noticeable rise in deaths primarily attributable to heart illnesses that has been occurring for many years. However, it can be proven that there are two intrinsic elements that could result in a misdiagnosis, namely the patient's understanding of the disease's severity and the attending physician's lack of familiarity with the factors.

[12]. Lack of an expert system increases the risk of misdiagnosis [Heart disorders, often known as cardiovascular diseases, are among the most fatal and dangerous diseases in the world. Heart rate and blood pressure (BP) vary from person to person, although the average person's pulse runs from 60 to 100 beats per minute, and the middle BP level may be around 120/80[15]. A number of catastrophic heart disease effects could arise from any restriction or irregularity in the heart's normal blood flow. They are among the deadliest conditions and are frequently referred to as cardiovascular diseases. Cardiovascular diseases include conditions affecting the heart, blood vessels, and the brain[17]. This paper deals with the analysis and impact of diabetes on patients suffering from Cardio Vascular Diseases and to measure accuracy using a prediction model.

This paper is organised as follows: Section 1 presents Introduction, Section 2 provides the literature survey pertaining to the study, Section 3 depicts the methodology, Section 4 presents Results and Discussion followed by Conclusion.

II. 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].

In this research, the authors compare various machine learning techniques, namely classification and forecasting algorithms, with regard to Cardiovascular Disease. The study uses machine learning techniques to conduct a thorough examination of about 41 papers on cardiovascular disease. This study thoroughly assesses the chosen papers and identifies gaps in the literature, allowing researchers to develop and apply their findings in clinical sectors, mainly on datasets pertaining to heart disease. The results of this study will help physicians anticipate potential heart dangers and take preventive action [3]. Different machine learning methods, including Support Machine Vector (SVM), K-Nearest Neighbor (KNN), Naive Bayes (NB), Artificial Neural Network (ANN), Random Forest (RF), and Gradient Descent Optimization, have been utilised by the authors in this paper (GDO). Intelligent Prediction of Cardiovascular Disease When compared to other classification algorithms, the Gradient Descent Optimization model delivers the best results. The GDO-based model's accuracy was 98.54% when its performance was evaluated. Additionally, 97.76% precision and 99.43% sensitivity (recall) have been noted. The system's prediction results are sufficient for its use in the diagnosis of cardiovascular disease [4].

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].

The authors of this study describe a real-time system for predicting heart illness based on Apache Spark, a powerful platform for large-scale distributed computing that can be utilized to successfully process streaming data events against machine learning using in-memory operations. The system comprises of two main sub parts: data storage and visualization and streaming processing [7]. The authors of this paper have proposed a machine learning-based prognostic modelling framework that can operate on static/low speed, streaming, and large amounts of data from wearables, such as fitness bands and biosensor watches, as well as static/high speed, large amounts of data from electronic health records. They employed a neuron network, which is a tool for producing incredibly precise findings in imprecise data. For two publicly available CVD data sets—the "NHANES" dataset and the "Framingham Heart Study" dataset—we have presented the results of the proposed framework implementation for static and low-velocity/volume settings from the EHR and clinical DBs, along with the experimental authentication of the planned framework[8]. In this study, researchers looked at how machine learning and artificial intelligence are used in the biomedical and healthcare industries. They give a quick literature review on how various data mining approaches can be used to forecast cardiac disease [9].

III. 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 parameter 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.

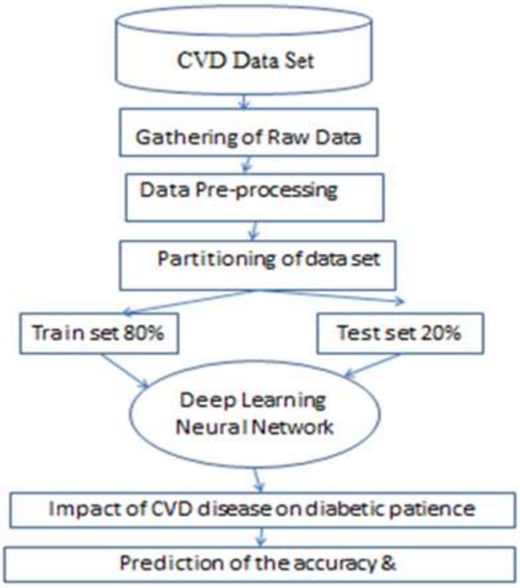

Fig.1 Block diagram of the proposed methodology

IV 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.

V Deep Neural Network

Deep neural networks (DNNs) are ANNs with more hidden layers between the input and output layers than ANNs, or additional depth. A more prominent machine learning technique is deep learning. It is used for more than only picture classification jobs; it also uses regular tabular data Learning the underlying structure of the data (input) vectors as well as the nonlinear mapping between the inputs and outputs is an interesting issue in deep learning. Training is the process of increasing a neural network's accuracy. A forward prop net's output is matched with the value that is known to be accurate. The discrepancy between the output that was generated and the output that was actually produced is the cost function or loss function.The discrepancy between the output that was generated and the output that was actually produced is the cost function or loss function Each edge and each node have their own weights and biases, which are used to change each set of inputs. The weights and biases of a neural network determine how accurate its predictions are. Training is the process of increasing a neural network's accuracy.. Compared to a conventional model, a deep neural network has more hidden layers [16].

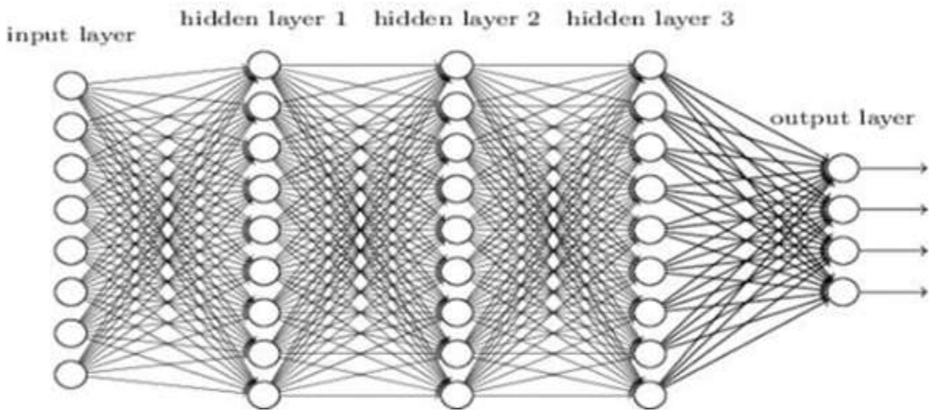

Fig. 2 Architecture of Deep Neural network

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.

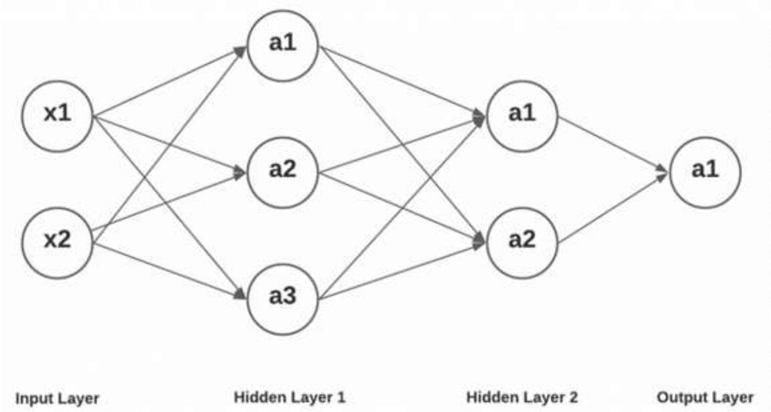

Fig 3 Block diagram of Deep Neural Network

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

$f(b + \sum_{i=1}^{n} x_i \, w_i)$ …. Eqn. (1)

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.

$f(z) = \sigma(z) = \frac{1}{1+e^{-z}}$ …. Eqn. (2)

Three output layers, six input layers, and 100 epochs were used in this study.

IV. Results and Discussion

Table-3 Significance of variable

Variable	Scaled Importance
HBA1C	100%
BUN	92%
BMI	71%
height	66%
CREA	60%

From the table-3 It can be shown that HBA1C has a 100% impact on the CVD model and is most significantly linked to the disease, which means that it is the variable that can best predict the CVD. Similar to BUN, the CVD model is affected by BUN to the tune of 92%. Height has a 66% impact on CVD, BMI (71% impact), and creatinine (60%) impact.

Table-4 Confusion matrix of CVD for Train Dataset

	Predicted		
Actuals		Bad	Good
	Bad	172	108
	Good	97	893

Table 4 Confusion matrix of CVD for Train Dataset is a representation of how a classification model performed; it provides details about both the actual and expected classifications made by a classifier. The accuracy was 83.85%, with true positives (893) and true negatives (172) False Negative (FN) = 97 and False Positive (FP) = 108 Sensitivity is equal to 100 TP/(TP+FN) = 100*893/990 = 90.20 percent. Specificity is calculated as follows: 100 TN/(FP+TN) = 100 172/280 = 61.42. The F1 score can be thought of as a weighted average of the precision and recall values, with the best and worst values falling at 1 and 0, respectively. 90% is the model's F1 score for the Train dataset

Table-5 Confusion matrix of CVD for Test Dataset

	Predicted		
Actuals		Bad	Good
	Bad	100	60
	Good	87	611

Table-5 depicts confusion matrix for the test dataset using CVD, It also goes by the name of an error matrix and enables model performance to be seen. It is a binary classification problem's measure, representing various pairings of actual vs. predicted values from table 5. True positive: The model accurately predicted 611 records of CVD patients. 60 entries contained false-positive predictions that they were CVD patients. 87 records were falsely forecasted as not having cardiovascular disease (CVD). True Negative: The model correctly predicted 100 records of non-CVD patients. One performance indicator, accuracy, is shown as a ratio of properly predicted observations to all observations. Accuracy is only tested when values of accurately predicted values are almost identical when we have symmetric datasets. The accuracy of this model is estimated to be 82.86%.

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%.

VI RoC for Training

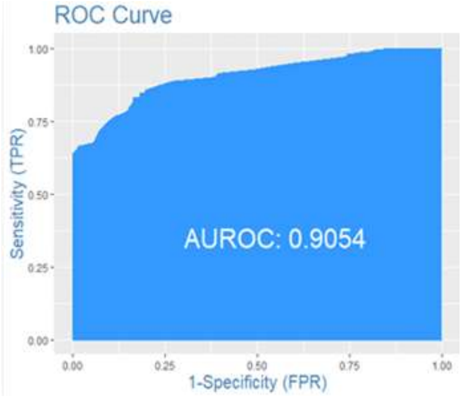

Fig. 4 : ROC curve - 1

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%.

VII ROC for Testing

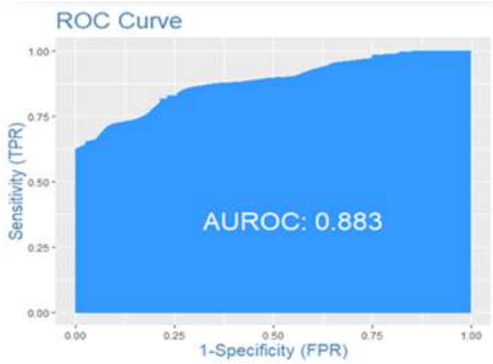

Fig. 5 : ROC curve - 1

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%.

VIII. CONCLUSION

In India, population growth and scarce resources coexist. As we have seen with Covid-19, there is a dire need for better healthcare. One of the important areas of the medical profession is the prediction of cardiovascular disease utilising the patient's existing data. There are numerous tools and strategies available for cardiovascular disease prediction. Deep Neural Network is employed in this study to categorise heart disease. Pre-processing activities like cleaning and identifying missing values are carried out to boost performance. The crucial component is feature selection, which improves algorithm accuracy and even focuses on the algorithm's behaviour. The suggested approach can aid in early patient diagnosis and can also help the healthcare industry forecast cardiovascular disease early in this concerning situation. The most accurate and effective methods for forecasting cardiovascular disease were found after a review of existing techniques. Due to the severity of cardiovascular illness, analysis of it is one of the most common fields of research in the modern era. This study demonstrated significant gains in performance for the identification of cardiovascular disease. This research has a greater impact because the accuracy and F1 score of the suggested methodology are clearly higher than those of the majority of the current methodologies.83.85%, 90%, and 90 % accuracy, F1score, and AUCROC have been recorded for train data set, similarly 82.86%, 89% and 88.3% for test dataset respectivelY.

Acknowledgements

The Bangalore-based Spectrum Diagnostic Centre & Health Care is duly acknowledged by the authors for providing support for research work by contributing relevant data.

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.

Dr. Manjula Sanjay Koti received her Ph.D, M.Phil, MCA from Bharatia University and Bapuji Institute of Engineering and Technology, Davangere in the year-2015, 2008 & 1999. She has around 25 publications including International Journals and International conferences along with a number of patents. She has also

sessioned various national & international conferences. Her areas of interest include Data warehousing. Data mining. Data Analytics. Pattern recognition and cloud computing.

B.H Alamma received M.Phil, MCA, B.E Electronics from Periyer University ,IGNOU and Bangalore University. Bangalore in the year 2006.2003,1991. Her area of interest includes Data warehousing , Data mining, Data analytics. Wireless networking. She has participated in conferences and seminars.