Study of Various Classification Approaches including Deep Learning in Heart Disease Prediction

Subhalaxmi Chakraborty, Department of Computer Science, University of Engineering & Management, Kolkata, India
Prayosi Paul, Department of Computer Science, University of Engineering & Management, Kolkata, India
Aditi Ghosh, Department of Computer Science, University of Engineering & Management, Kolkata, India
Suparna Bhattacharjee, Department of Computer Science, University of Engineering & Management, Kolkata, India
Soumadeep Sarkar, Department of Computer Science, University of Engineering & Management, Kolkata, India

***Abstract*—**During the last couple of decades, apart from other diseases, heart disease has proved to be a major cause of human death. This refers to a wide variety of heart conditions which include diseased vessels, structural problems and blood clots. This paper is aimed to make an effective prediction about the vulnerability to heart disease of a person depending on the given health parameters at a very early stage & to reduce premature death, with an improved accuracy and reliability compared to other traditional models. Our proposed prediction model has proved to be reliable in this connection and has yielded a maximum accuracy of 99.0% using Random Forest, Support vector machine and Decision trees and K-nearest neighbour algorithms. Using Cross Validation, we have also prepared the model (with the highest accuracy level of 98.6%) to work efficiently, taking the correct pattern of the dataset. Further, the proposed model has outperformed other traditional models in terms of accuracy using some other algorithms (Hybrid Ensemble model, Extreme Gradient Boosting with Random Forest and Stochastic Gradient Descent with accuracies of 94.2%, 87%, 78% respectively). Moreover, TensorFlow has also been used in order to get reliable prediction of heart disease with an approximate accuracy of 83.44%. The novelty of our proposed model lies in showcasing better results compared to those obtained by the traditional models and this model can easily be applied in medical science to provide better diagnosis.

Keywords—Heart disease, Random Forest, Support vector machine, Decision trees, KNN, TensorFlow.

I. INTRODUCTION

Heart disease is a lethal cause that is responsible for sudden death. People all over the world are suffering from cardiovascular diseases. Every year approximately 17.9 million people die from cardiovascular disease according to World Health Organization [1]. Today, everyone is familiar with these two common terms: heart attack and stroke. Heart disease encloses a vast variety of symptoms and causes which are categorized into different terms like Coronary Artery Disease (CAD), Cardiomyopathy, Atherosclerosis [2] with some common symptoms like chest pain, shortness of breath, nausea. During the last couple of decades with the growth of modern technologies, medical surgeries and treatments also reached a high level of success. Approximately, in India, out of every 100 people 23 die from heart disease [3]. According to the European Society of Cardiology, 26 million people of heart disease were diagnosed and every year near about 3.6 million people suffered from heart disease [4]. From literature [5], it is observed that the scarcity and absence of skilled and experienced medical experts and doctors in some countries is also a great problem. In such scenario heart disease prediction using machine learning has created a new approach to reduce the premature death. Machine learning algorithm models can forecast the onset of disease and has played significant role in data processing and investigation. Accuracy of any model depends on the working perfection of the algorithm. Machine Learning has several standard algorithm techniques which includes Random Forest, Decision Tree, SVM, K Nearest Neighbour, Logistic Regression.

Using Machine Learning Algorithms, Kohli et al. [6] have worked on prediction of heart disease using linear regression which has given the accuracy of 87.1%. Palaniappan et al. [7] have presented an Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques such as Neural Networks (NN), Naive Bayes (NB), Decision Tree (DT) and concluded that NB model gives the best correct prediction among the others which is 86.12%, NN model gives 85.68% and DT model achieved 80.4% correct prediction.

Gudhade et al. [8] has realized a decision support system based on MLP neural network and Support Vector Machine (SVM) architecture for the classification of heart disease. Using the SVM approach it has been concluded with an accuracy of 80.41% by them. They have used the Artificial Neural Network which classifies their data with 97.5% accuracy into 5 categories.

Cheng et al. [9] have worked to build an architecture to evaluate Carotid Artery Stenting prognosis by using Artificial Neural Network and they showed that the performance of ANN model with an accuracy of 82.5% in testing set and 80.76% accuracy in overall patients. For identifying heart disease, Das et al. [10] have presented a procedure which have used a SAS base software 9.1.3 and in the centre of the proposed system there is a Neural network (NN) ensemble method. They worked for this methodology by taking data from Cleveland heart disease database and have acquired a classification accuracy of

89.01%. In this paper the standard algorithms of machine learning and their implementation on heart disease detection dataset is described along with their individual significance. The motive of this paper was to create such a model which will be able to give the most accurate result using the standard algorithms along with hybrid ensemble, cross validation, gradient descent, XGBRF and CNN using TensorFlow [11]. This model is aimed at detecting a person's sensitivity to heart disease using Machine Learning Algorithms from different aspects along with Hybrid Ensemble Model, Extreme Gradient Boosting with Random Forest and Stochastic Gradient Descent. Here the dataset of heart disease obtained from Kaggle is used to address the binary classification problem of heart disease based on parameters namely age, pulse rate, cholesterol, blood pressure, blood sugar and many more parameters are there. The models had predicted the result based on the train and test data which is divided in a 70-30 ratio.

II. PROBLEM DESCRIPTION

The dataset contains factors which have an intense effect on the condition of human heart. In this paper, the parameters considered in order to compute the major risk percentage according to CAD [12] are: age, sex, blood pressure, heart rate, diabetics, cholesterol. Moreover, some extra features are also added in the dataset like: Chest pain (Cp), Fasting blood sugar (fbs), Rest ecg, Maximum heart rate achieved (thalach), Exercise induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), the slope of the peak exercise ST segment (slope), Target. Fig. 1 represents the correlation between the various attributes. The flowchart of the proposed model is represented by Fig. 2.

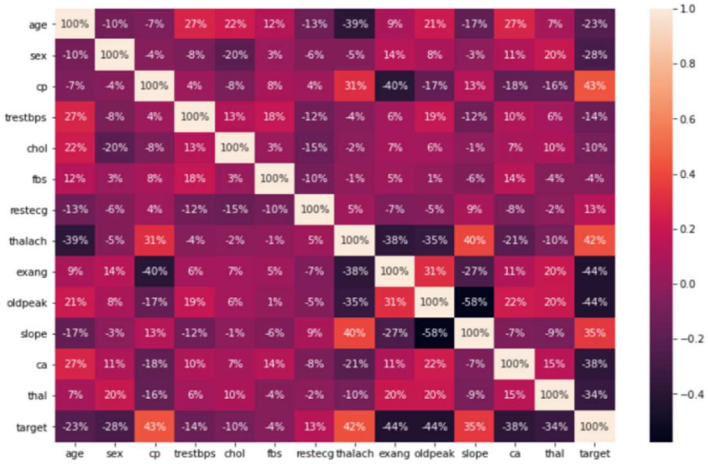

Fig. 1: Heatmap representation of correlations between attributes

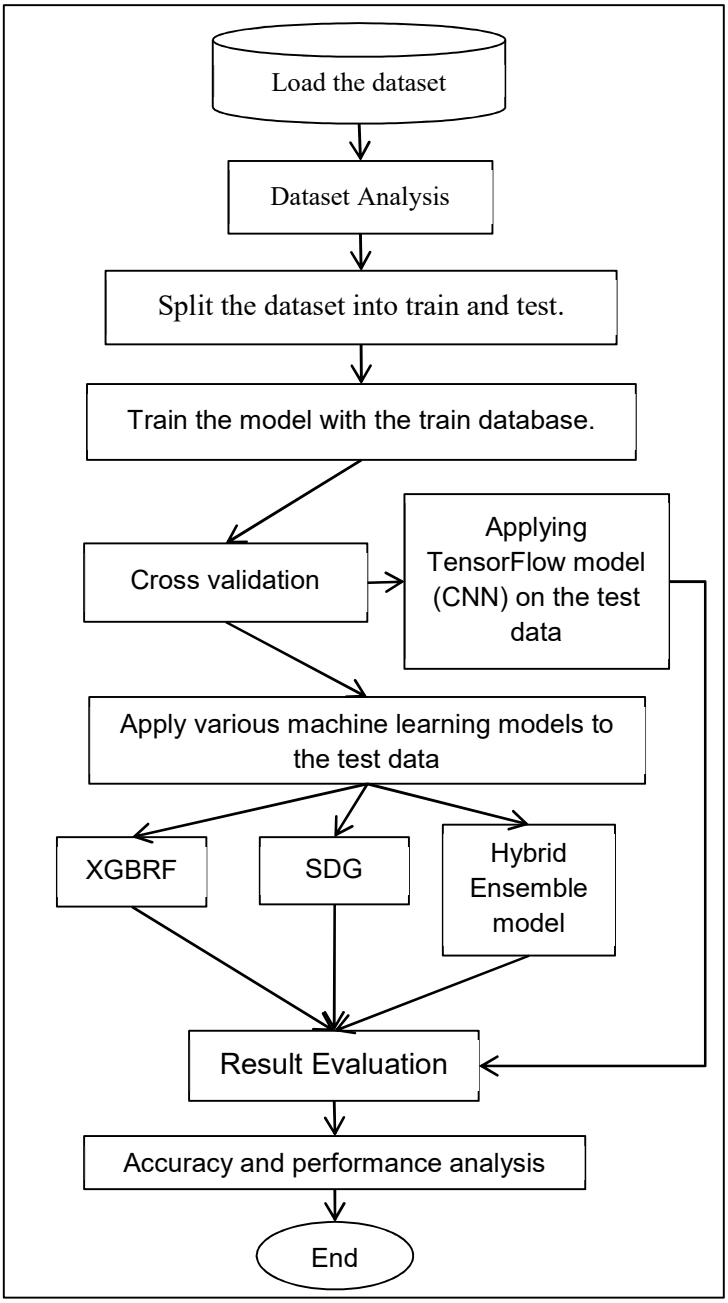

Fig. 2: Flowchart of the proposed model.

III. OVERVIEW OF THE CLASSIFICATION METHODS AND THEIR IMPLEMENTATION:

Stochastic Gradient Descent

Stochastic Gradient Descent is a well-known optimization approach for finding the model parameters that tie in with the best fit between predicted and actual outputs. In machine learning's application field, the usage of stochastic gradient descent is increasing day by day. It favours to optimize

different problems in deep learning and also can be considered as a faster alternative for training support vector machines.

For training deep learning model, sometimes we consider the objective function as a sum of a finite number of functions:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \cdots\cdots\cdots (i)$$

where n is the size of training dataset, i is training data instance index, $f_i(x)$ is a loss function based on the training data instance indexed by i. Graphical representation of stochastic gradient descent is represented by Fig. 3.

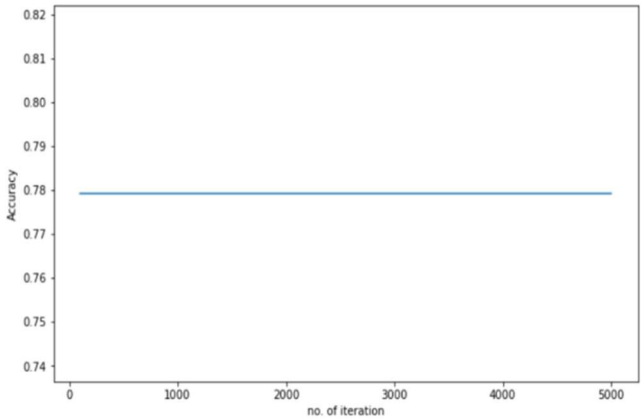

Fig. 3: Graphical representation of Stochastic Gradient Descent

Random Forest

This classifier is used to construct several decision trees during the training of that model. After that being an "Ensemble technique", it combines all the predictions gained from all the trees and makes the final one. In this case, the importance of the attribute is based on the decrease of node impurity calculated by the probability of reaching that node where probability = (no. of samples reach the node/total number of samples) [13].

The importance for each feature is calculated in Scikit-Learn as

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k} \quad \text{------- (ii)}$$

where,
fi sub(i) = the importance of feature i,
ni sub(j) = the importance of node j

It can be normalized as:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j} \quad \text{------- (iii)}$$

Now the final feature importance:

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T} \quad \text{---------- (iv)}$$

where RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model, normfi sub(ij)= the normalized feature importance for i in tree j, T =total number of trees.

K Nearest Neighbour

The main objective of KNN classifier is to predict the class of a given data point by identifying the class of the nearest observation. The scale of the variables matters for the prediction. As any variable present on large proportion will have a greater impact on the distance between the observations, and hence on the KNN classifier, than variables that are on a small proportion [14]. In this procedure "K=1" is taken first to get the prediction. In order to observe if a better result is possible, an error rate versus K value plotted. The K value for least error is considered for best result. This is the plotted graph in this research paper which gives the best K value as 1. Here, Fig. 4 refers to the error representation of KNN.

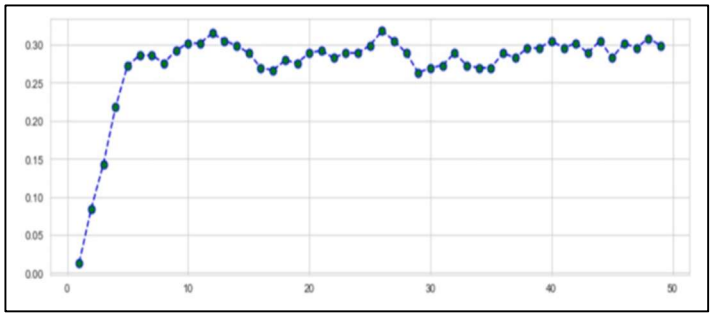

Fig. 4. Graphical representation of error calculation of KNN

Support Vector Machine

It is one of the supervised learning algorithms which is used to create the best decision boundary segregating n-dimensional space into different classes and the best decision boundary is called "Hyperplane". So, it chooses extreme points ("Support Vectors") to make the hyperplane [15]. Here, we have made changes on some parameters, such as-setting "C" value in increasing order to make the hyperplane much smoother, setting "Gamma" value in decreasing order to make hyperplane tuning and setting "kernel" as "rbf", we have tried to find out the best hyperplane for our model.

Naive Bayes

It is also one of the most efficient Supervised learning algorithms which is basically based on "Bayes Theorem". Naive Bayes classifier needs a small training data to determine the parameters needed for classification and it assumes that the value of a particular feature is independent of the value of any other features [16]. So, while working with continuous data, an assumption often taken is that the continuous data corelated with each class are distributed according to a normal (or Gaussian) distribution. This feature will be:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad \text{------- (v)}$$

Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution.

Logistic Regression

In linear regression some estimated probability may be negative in order to balance and make it close to zero. For a binary classification such a negative result does not make any sense. Here the logistic regression is introduced and it uses a logistic function [14].

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \quad \text{------- (vi)}$$

Here p(X) is the probability of a dataset named default. This expression ensures that the output will always lie between 0 and 1 for all values of X where X is the set of predictors. It is noticed that for low balances the prediction is close to, but not less than zero. Likewise, for high balances prediction is close to, but not more than one. The logistic function will always produce a sigmoid curve shown in figure 5, and not depended on the value of X, a sensible prediction is achieved. The curve is as follow-

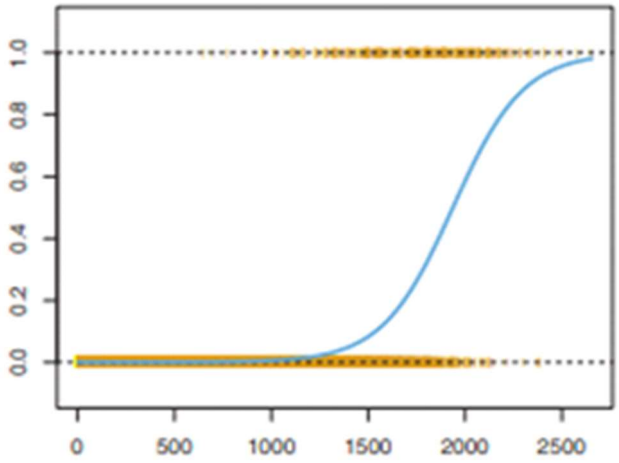

Fig. 5: Sigmoid curve of Logistic Regression

Decision Tree

Using decision trees, it is predicted that each observation belongs to the most commonly occurring class of training observations in the area to which it belongs. To analyse the results of a tree, it is better to show interest not only in the class prediction corresponding to a particular terminal node region, but also in the class proportions among the training observations under that region [13].

The associated term is entropy which is generally used to evaluate the quality of a particular split. Entropy can be represented by the following expression:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \quad \text{--- (vii)}$$

Here, $\hat{p}_{mk}$ represents the proportion of training observations in the m^{th} region that are from the k^{th} class.

Cross Validation

Cross validation is an important technique which is used to assure whether our model is getting the correct pattern of the dataset without getting too much noise. In this process, the model is trained using a specific part of the dataset and then trained using the complementary subset of the data-set. Here, we have used "K-Fold Cross validation", in which method the dataset is divided into "k" number of subsets & then training is performed on all the subsets leaving one (k-1) subset for the evaluation of the trained model. Thus, the iteration is done "k"(taken k=10) times with a different subset reserved for

testing purposes each time. It is simpler to examine the detailed results of the testing process.

```
Accuracy: 0.9860529986052998
Sensitivity: 0.9804469273743017
Specificity: 0.9916434540389972
```

Fig. 6: Graphical Representation of Accuracy, Sensitivity and Specificity after Cross Validation

Extreme Gradient Boosting with Random Forest

Extreme Gradient Boosting is used to introduce the techniques to speed up the training of the model and to make better result overall performance of the model. Here, we have configured this with an ensemble tree algorithm (Random Forest) where a random portion of the input variables in the tree at each split point is considered. This guarantees that each tree included in the ensemble is effective, but different in random ways. We have done this as follows:

```
[[114  26]
 [ 15 153]]
```

```
[46] print(classification_report(y_test,p1))
```

	precision	recall	f1-score	support
0	0.88	0.81	0.85	140
1	0.85	0.91	0.88	168
accuracy			0.87	308
macro avg	0.87	0.86	0.86	308
weighted avg	0.87	0.87	0.87	308

Hybrid Ensemble

The objective is to improve the performance results of machine learning problems for multiclass classification problems using this algorithm by using several classifiers in an ensemble. In this paper this approach has been used to get a more appropriate accuracy [14].

In this method more than two separate models are created with the same dataset. A new ensemble model is created based on voting and aggregation of the results of their performance helps in the ultimate evaluation of this model. Here in this paper six methods are used in order to get the ensemble model. Those are logistic regression, naïve bayes, random forest, decision tree, k nearest neighbour and SVM.

	precision	recall	f1-score	support
0	0.95	0.92	0.93	140
1	0.94	0.96	0.95	168
accuracy			0.94	308
macro avg	0.94	0.94	0.94	308
weighted avg	0.94	0.94	0.94	308

TABLE-I: Experimental Evaluation and Corresponding Result along with the Comparative Discussion [17].

Model	Accuracy	Sensitivity	Specificity	FP rate
RF	0.986	0.980	0.992	0.0178
DT	0.975	0.975	0.975	0.2020
NB	0.809	0.844	0.774	0.1130
SVM	0.972	0.958	0.986	0.2941
KNN	0.971	0.958	0.983	0.2023
LR	0.844	0.866	0.822	0.0595
HEM	0.931	0.958	0.921	0.0416

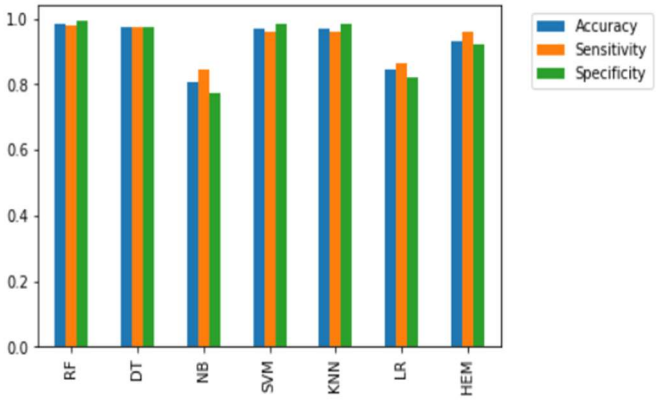

Fig. 6: Graphical Representation of Accuracy, Sensitivity and Specificity after Cross Validation of various models mentioned in TABLE-I.

TABLE-II: Experimental Evaluation and Corresponding Result along with the Comparative Discussion

Model	Accuracy	Sensitivity	Specificity
GD	0.779	0.714	0.857
HEM	0.942	0.958	0.921
XGBRF	0.867	0.911	0.814

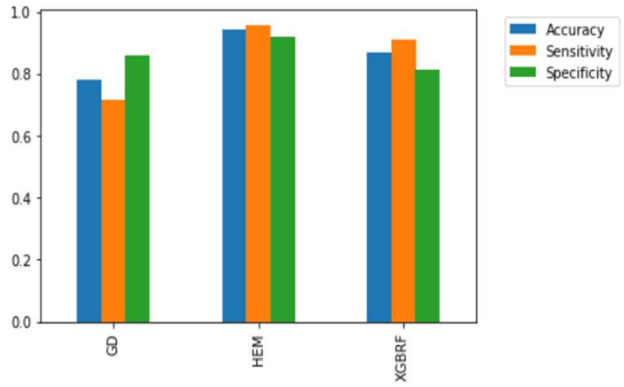

Fig. 7: Graphical Representation of Accuracy, Sensitivity and Specificity for GD, HEM and XGBRF mentioned in TABLE-II.

Using TensorFlow (CNN Or Convolutional Neural Network)

Being the heart of deep learning, a neural network contains multiple perceptron layers where perceptron is the basic building block of network. To build a network of perceptrons, we can connect layers of perceptrons, using a multi-layer perceptron model. Every neural network contains three portions and those are: 1) input layer, 2) output layer, 3) hidden layer. CNN is the version of a neural network where the concept of multilayer perceptron is used. It is more often used in solving classification problems and analysing visual data. This CNN model consists of three layers and those are: 1) convolutional layer, 2) pooling layer, 3) fully-connected layer. It is considered as more advantageous than others as the pre-processing required in a CNN is much less compared to other classification algorithms. While in primitive methods filters are needed to give manually, with enough training, CNN have the ability to learn these characteristics. The role of the CNN is to bring down the data into a suitable form which is easy to process, without losing crucial attributes which are necessary for getting a good prediction result. Here, model accuracy and model loss are represented by Fig. 8 and Fig. 9 respectively. It is observed that as the number of epochs is increasing from 0 to 100, the model accuracy is increasing and the model loss is decreasing gradually.

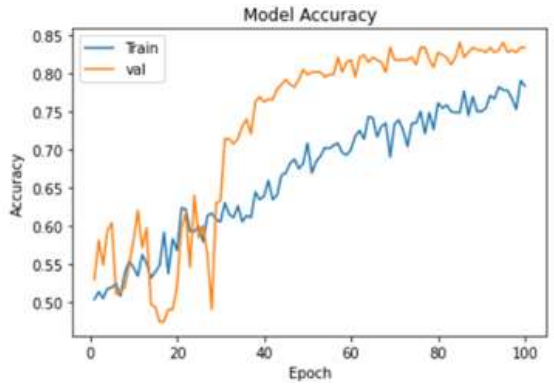

Fig. 8: Graphical representation of Model Accuracy

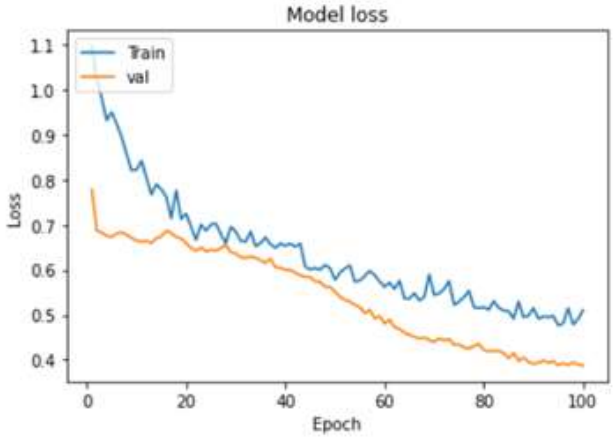

Fig. 9: Graphical representation of Model loss

IV. CONCLUSION AND FUTURE WORK

In order to process the unorganized raw data related to heart disease the Machine Learning approach is used and a better accuracy has been gained. In this research paper the main objective is to provide a new and novel discernment towards heart disease. To accomplish this challenging objective along with various classic machine learning algorithms (RF, XGBRF, SGD, KNN, DT, SVM, HEM, LR, etc.), CNN is also used and which is implemented using Tensor flow. Due to its reliability & accuracy this model can be adopted as the basic treatment aid to detect heart disease at a very early stage which can drastically control the mortality rate. Machine learning is a growing field of various invented technologies as well as opportunities. The idea of innovation is not bound under any limit so further extension of this paper is attainable. New feature selection technique as well as data processing methods can be included and better accuracy can be gained.

REFERENCES

[1] "Cardiovascular diseases (CVDs)", World Health Organisation (WHO), https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), Accessed 30th July, 2021.

[2] "Heart disease", Mayo Clinic, https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118, Accessed 25th June, 2021

[3] S. Shylaja, & Muralidharan S. 2R., "Hybrid SVM-ANN Classifier is used for Heart Disease Prediction System", International Journal of Engineering Development and Research, vol. 7, no. 3, pp. 365-372, 2019.

[4] J. Lopez-Sendon, "The heart failure epidemic," Medicographia, vol. 33, no. 4, pp. 363–369, 2011.

[5] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart disease diagnosis," J. Intell. Learn. Syst. Appl., vol. 5, no. 3, 2013.

[6] Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Diseases Prediction," 4th International Conference on Computing Communication and Automation (ICCCA), 2018.

[7] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in IEEE/ACS International Conference on Computer Systems and Applications. IEEE, pp. 108–115, 2008.

[8] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision Support System for Heart Disease based on Support Vector Machine and Artificial Neural Network," Computer and Communication Technology (ICCCT), International Conference, pp. 741–745, 2010.

[9] C.-A. Cheng and H.-W. Chiu, "An Artificial Neural Network Model for the Evaluation of Carotid Artery Stenting Prognosis Using a National-Wide Database," Proc. 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2566–2569, 2017.

[10] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," Expert Syst. Appl., vol. 36, no. 4, pp. 7675–7680, May 2009.

[11] Jian Ping Li, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, and Abdus Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," IEEE Access, Vol. 8, 2020.

[12] National Health Council,' Heart Health Screenings', 2021. [Online] Available: https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/heart-health-screenings

[13] Scott Hartshorn, "Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners Kindle Edition", 2016.

[14] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, "An Introduction to Statistical Learning with Applications in R", 2017.

[15] Nello Cristianini, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods Illustrated Edition, Kindle Edition", 2000.

[16] Shadab Adam Pattekari and Asma Parveen, "Prediction System for Heart Disease Using Naive Bayes," International Journal of Advanced Computer and Mathematical Sciences, Vol. 3, Issue 3, 2012, pp. 290-294, 2012.

[17] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, "Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data," Sensors, vol. 20, no. 9, p. 2649, May 2020.