Cardiac Arrest Prediction using Machine Learning Algorithms

Utsav Chauhan
HMR Institute of Technology and Management
Delhi,India
utsavchauhan07@gmail.com

Vikas Kumar
HMR Institute of Technology and Management
Delhi,India
chauhanvicky577@gmail.com

Vipul Chauhan
HMR Institute of Technology and Management
Delhi,India
vipulrajput01@gmail.com

Sumit Tiwary
HMR Institute of Technology and Management
Delhi,India
sumitsahil.chapra@gmail.com

Amit Kumar
Assistant Professor, CSE Dept.
HMR Institute of Technology and Management
Delhi,India
amitkr6002@gmail.com

***Abstract* - In today's era cardiac arrest and other heart diseases are the most common problem in majority of people, and there are various factors that act as backbone of this problem like people are not paying attention towards health mainly because of work stress, laziness, substandard quality of food that results in increasing cholesterol and untimely diagnosis of heart disease due to lack of technology, methods used in diagnosing these diseases and consequently having a lot of tests. A lot of research and medical supporting systems are developing day by day, however, every system have its various features or advantages and limitations which are unknown to either side. This paper aims to study different machine learning algorithms on dataset to predict the possibility of a cardiac arrest based on various controlled and uncontrolled variables.**

Keywords- Cardiac arrest;diagnosis;medical supporting systems; machine learning algorithms

I. INTRODUCTION

Human body consists of organs that are very prone to disease like heart, brain, kidney, liver, etc. One of the most crucial and disease prone is heart inspite of having a normal lifestyle; tendency of having cardiovascular disease due to blocked blood vessel resulting in chest pain,stroke and other symptoms.[1][2][3] There are a lot of factors responsible for heart disease that can or cannot be controlled. For instance: age, gender, family history cannot be controlled whereas obesity, physical activity, diet can be controlled. Now a days, some doctors use machine learning system to detect heart disease but some don't. According to WHO, cardiovascular disease(CVD) is on the top of cause to death globally and people suffering from CVD or with high risk need to be diagnosed early in order to tackle the problem.[4]

Artificial intelligence is one of the technology that is solving this problem by predicting the risk percentage using Machine Learning. In Machine Learning, several predictive an classification algorithms such as Random Forest, decision tree, Linear Regression, Support Vector Machines (SVM) and Artificial Neural Networks (ANN)[5]. ANN is the most powerful tool in artificial intelligence due to its brain like functioning in which a neuron (also known as perceptron) is the fundamental unit of the network and a number of inputs are fed on one neuron which gives an output.

If by using a certain algorithm we could predict the occurrence of the cardiac arrest, then that would prove as a milestone in the field of both engineering and medical sciences.[11] Not only would it count as an advancement but it could help prevent the death by cardiac arrest up to a significant level and that would directly affect the life expectancy rate.

The remainder of the paper is organised as follows. Section II of the paper presents the literature review and Section III describves the parameters used. Section IV of the paper presents the algorithms used including Support Vector Machine, Random Forest, Decision Tree, Logistic Regression and Artificial Neural Network.

II. LITERATURE SURVEY

There have been many attempts at predicting cardiac arrests through various algorithms such as Linear Regression. [6] Various permutations of algorithms have been employed to increase the accuracy of the prediction outcome so that it may become reliable. [7] Even data analyzing techniques such as Neural Networks, Naïve Bayes and Decision tree have been employed by Pattekari et al. to create a system for predicting Cardiac Arrests, though it only discussed the theoretical model of the system.[8] Even through IoT, cardiac arrest prediction have been tried, which focused mainly on the people belonging to an old age.[9]

III. PARAMETERS USED

The parameters used while implementing the algorithm were age, sex, chestpain, restbp, chol, fbs, RestECG, MaxHR, ExAng, Oldpeak and Decision.

While tidying the data ,the dataset used contained some missing values which had to be removed before the algorithms could be applied on it.The missing values in the dataset were removed by filling the spots with the mean values of the column. The dataset was checked for the outliers and none were found while mapping the data.

IV. ALGORITHMS USED

Several algorithms such as Support Vector Machine, Random Forest, Decision Tree, Logistic Regression and Artificial Neural Network are detailed in the section below.

A. Support Vector Machine:

Support vector machine is a classification algorithm. It is a supervised machine learning algorithm. It is preferable when the data is of high dimension. One of the main reasons to use the algorithm was because of the kernels. While implementing SVM, different Kernel function can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels, which makes it more versatile. In SVM, the data is mapped to a high dimension space which categorizes the data even if the data is linearly inseparable. A separator between the categories is found after which the hyperplane is drawn considering the separator. After this, when the next set of input is fed the algorithm can predict in which class the inputs will lie. The support vectors are observation points whereas the SVM is a front which differentiates the two classes.[9][10]

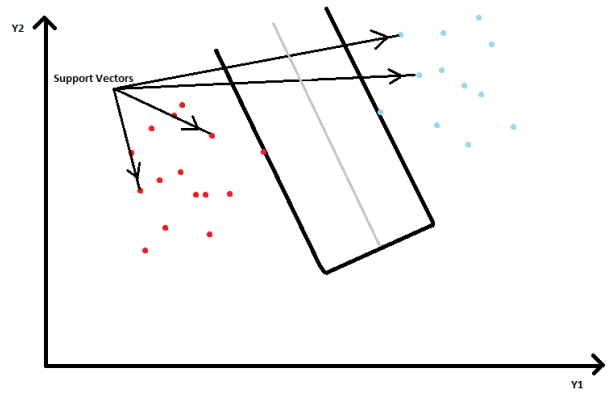

Fig 1. Example of Support Vector Machine Algorithm

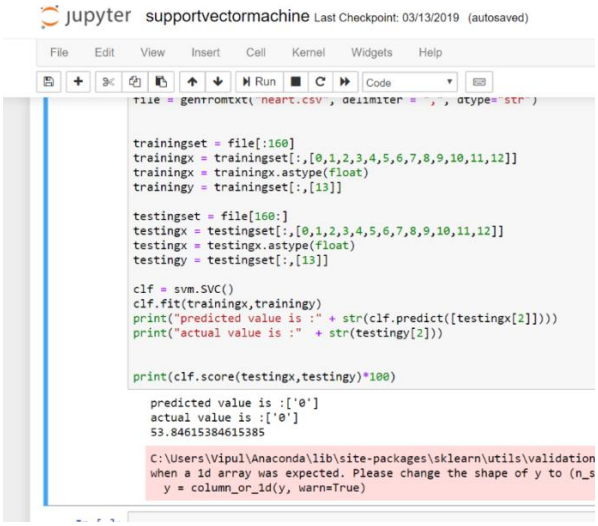

Fig. 2. Accuracy achieved (*using SVM: 53.8462%)*

B. Random Forest :

Random forest is flexible algorithm used in machine learning, frequently used due to its simplicity and the advantage of being able to use in regression and classification [11]. Forest is the ensemble of multiple decision tree by adding extra randomness to the model during the growth of trees and taking some random subset of features are taken into consideration. Random Forest searches for the best feature among the random subset resulting in wide diversity of outcomes i.e. better model. Sklearn provides afeature of measuring the relative significance of features. The problem in regression arises is overfitting which can be reduced using this algorithm as averaging of several decision trees are done lowering the risk of overfitting. Another regression problem is high variance which is not prominent in random forest by using multiple trees, the probability of stumbling across a classifier that does not perform well due to relationship between train and test data.

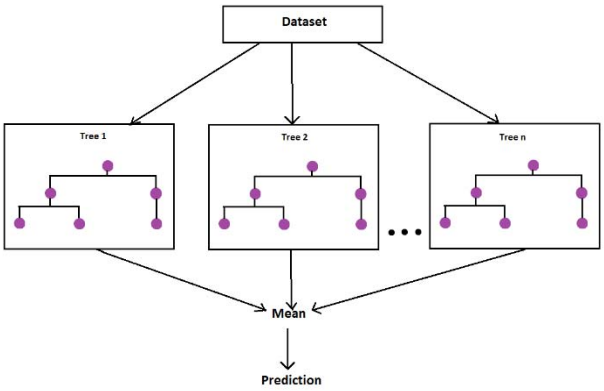

Fig. 3.Example of Random Forest Algorithm

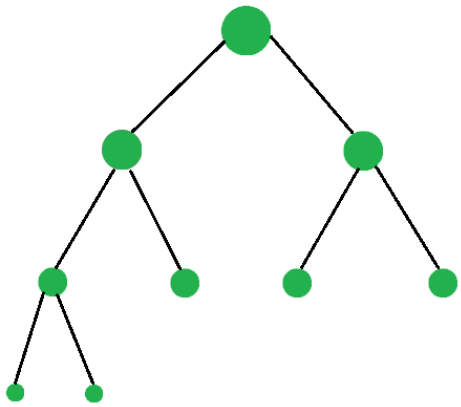

Fig. 5. Example of Decision Tree Algorithm

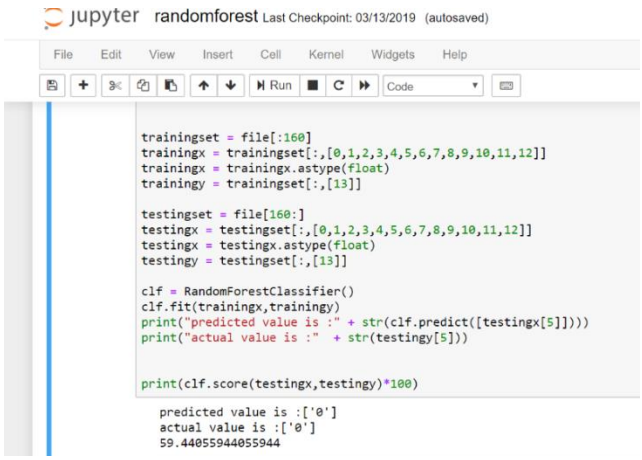

Fig. 4. Accuracy achieved (*Random Forest : 59.4405%*)

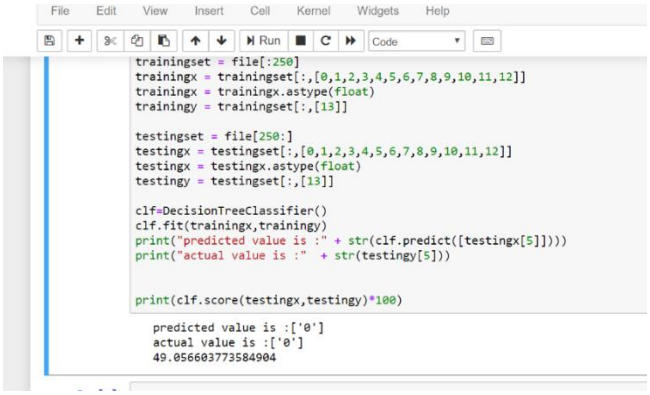

Fig. 6. Accuracy Achieved (*DecisionTree : 49.0566%*)

C. Decision Tree

Decision tree algorithm is a supervised learning algorithm that is developed to solve the problems of regression and classification.[12] So, the main advantage of decision trees is that they can handle both numerical and categorical data. Like other conventional algorithms decision tree algorithm creates a training model and that training model is used to predict the value or class of the target label/variable but here this is done by learning decision rules inferred from previous training dataset. This algorithm makes use of tree structure in which the internal nodes also known as decision node refers to an attribute and each internal node has two or more leaf nodes which corresponds to a class label. The topmost node known as root node corresponds to the best predictor i.e. best attribute of the dataset. This algorithm splits the whole data-frame into parts or subsets and simultaneously a decision tree is developed and the end result of this is a tree with leaf nodes, internal nodes and a root node. As the tree becomes more deep and more complex, then the model becomes more and more fit.

D. Logistic Regression

Logistic Regression is a type of regression in which ‘y’ is the target variable (binary) predicted on input variable ‘x’ comprising linear weight or values. In this case, the target variable is ‘Decision’ which decides whether a person having a chance of cardiac arrest and they should consult a doctor or have normal conditions. The best fit curve is obtained so that according to that curve, target variable can be predicted based on variable ‘x’ comprising distinct parameters. Probability function is the function obtained after the training of data set that is later transformed to binary values (0, 1) for the real probability prediction. One of the major reasons of using it is that we can include more than one explanatory variable (dependent variable) and those can either be dichotomous, ordinal, or continuous. Also logistic regression provides a quantified value for the strength of the association adjusting for other variables.

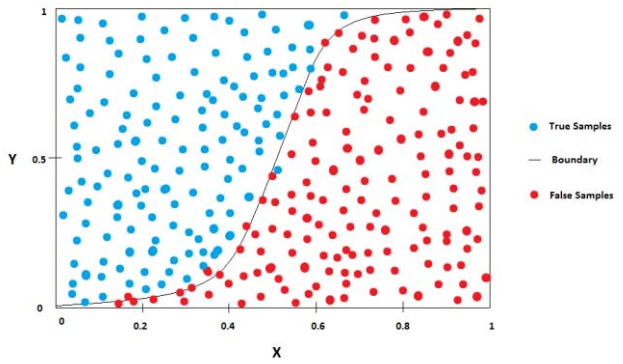

Fig. 7. Example of Logistic Regression Algorithm

```
trainingset = file[:200]
trainingx = trainingset[:,[0,1,2,3,4,5,6,7,8,9,10,11,12]]
trainingx = trainingx.astype(float)
trainingy = trainingset[:,[13]]

testingset = file[200:]
testingx = testingset[:,[0,1,2,3,4,5,6,7,8,9,10,11,12]]
testingx = testingx.astype(float)
testingy = testingset[:,[13]]

lr = linear_model.LogisticRegression()
lr.fit(trainingx,trainingy)
print("predicted value is :" + str(lr.predict([testingx[5]])))
print("actual value is :"  + str(testingy[5]))


print(lr.score(testingx,testingy)*100)
```

```
predicted value is :['1']
actual value is :['3']
56.310679611650485
```

Fig.8. Accuracy achieved (*Logistic Regression : 56.3106%*)

E. Artificial Neural Network

An artificial neural network is a collection of a number of neurons in a structure similar to that of a human brain and based on the functions of biological neural networks. The structure of a neuron consists of three parts the center is known as "nucleus", the tail is known as "axon" and the head of neurons consists of a number of branches called "Dendrites". A single neuron is not that strong but when we have a lots of neurons together they work together to do magic. Based on this principle artificial neural networks are used to predict whether the person may have a cardiac arrest or not. The ANN consists of majorly three parts the input layer, the hidden layers and the output layer. The input layer consists of number of neurons equals to the attributes or say independent variables that are used to predict the value of dependent variable or target label which corresponds to output layer. The weighted output from the input layer neurons are fed to the hidden layer neurons and here in hidden layer neurons activation function is applied. Activation functions are very important for a neural network as it helps the network to learn and introduce non-linear properties to the network, with certain mathematical computations, an activation function coverts a input signal of a neuron to an output signal [13-16]. Here "ReLU" activation function is used in order to predict the target label and the main purpose of using ReLU (Rectified Linear Unit) as an activation function is that the function and its derivative both are monotonic [17]. In this model, a single hidden layer consisting of eleven neurons is used on which activation function "ReLU" is applied with a learning rate of 0.01, solver "adam" and number of epochs are taken 100. And using these hyper parameters artificial neural network is trained.

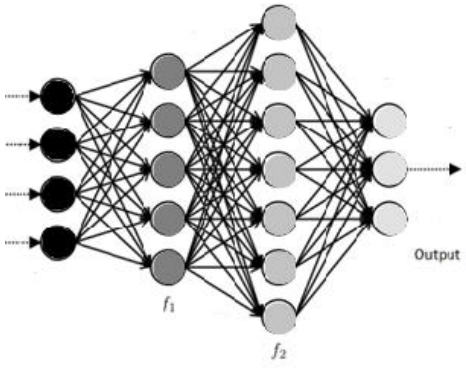

Fig. 9. Example of Artificial Neural Network

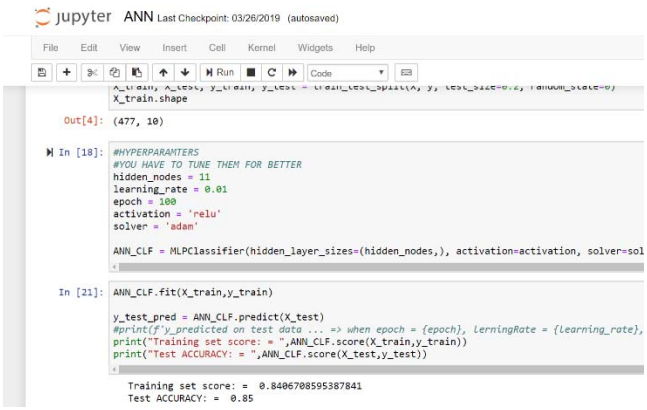

Fig. 10. Accuracy achieved (*ANN : ~85.00%*)

TABLE 1. ACCURACY TABLE

S.NO	ALGORITHM	ACCURACY
1	Support Vector Machine	53.8462%
2	Random Forest	59.4405%
3	Decision Tree	49.0566%
4	Logistic Regression	56.3106%
5	Artificial Neural Network	85.00%

The accuracy was found highest while implementing the neural network. Since the neural network learns after every iteration, therefore the mean value the accuracy results were taken and used as the final accuracy of ANN.

V. CURRENT PROBLEM AND THE PROPOSED SOLUTION

The only shortcoming faced during the implementation of the algorithm to predict the cardiac arrest is the unavailability of the larger dataset than used [13]. If a larger dataset was available then the neural network would have been trained more accurately and the results would have been more precise than they presently are.

When a much larger dataset is available then ANN could be applied and also the epochs and hidden layers could be increased which would increase the accuracy and precision of the outcome provided by the neural network.

VI. CONCLUSION

After applying Support Vector Machine, Random Forest, Decision Tree, Logistic Regression and Artificial Neural Network algorithms on the dataset to predict the occurrence of cardiac arrest in patients it is found out that the accuracy of the Artificial Neural Network is the highest (~85 %). Also, since the dataset was limited, the accuracy of the algorithm is low. Had the dataset been bigger the accuracy of ANN would also have been much higher than the current average value of the outcome. Thus Artificial Neural Network can be used as a basis to predict occurrence of cardiac arrest in the patients according to the results of the algorithms performed on the dataset.

VII. REFERENCES

[1] Rezuş, C., Moga, V. D., Ouatu, A., & Floria, M. (2015). QT interval variations and mortality risk: Is there any relationship?. Anatolian journal of cardiology, 15(3), 255..

[2] Wellens, H. J., Schwartz, P. J., Lindemans, F. W., Buxton, A. E., Goldberger, J. J., Hohnloser, S. H., ... & Myerburg, R. J. (2014). Risk stratification for sudden cardiac death: current status and challenges for the future. European heart journal, 35(25), 1642-1651.

[3] Nishibe, T., Sato, K., Yoshino, K., Seki, R., Yana, K., & Ono, T. (2012, August). RR-QT interval trend covariability for sudden cardiac death risk stratification. In 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 4287-4290). IEEE.

[4] E. Ebrahimzadeh, M. Pooyan, and A. Bijar, "A novel approach to predict sudden cardiac death (SCD) using nonlinear and time-frequency analysis from HRV signals," PLoS ONE, vol. 9, no. 2, 2014.

[5] Vahid Houshyarifar and Mehdi Chehel Amirani, "An approach to predict sudden cardiac death (SCD) using time domain and bispectrum features from HRV signal," Bio-Medical Materials and Engineering, vol. 27, no. 2–3, pp. 275–285, 2016.

[6] C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, vol.2, 1998, pp.121- 167

[7] Deep Bera and Mithun Manjnath Nayak , Mortality risk assessment for ICU patients using logistic regression,bComputing in cardiology , 2012.

[8] S. A. Pattekari and M. A. Yadav, "Heart attack prediction system using data mining techniques." International Journal of Ethics in Engineering & Management Education, vol. 1, no. 1, pp. 34–37, Jan. 2014.

[9] J. -V. Lee and Y. -D. C. a. K. T. Chieng, "Smart Elderly Home Monitoring System with an Android Phone," International Journal of Smart Home, vol. 7, pp. 17-32, May 2013.

[10] Fei, Ye. "Simultaneous Support Vector selection and parameter optimization usingSupport Vector Machines for sentiment classification." Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on. IEEE, 2016.

[11] Y Wei, T Liu, R Valdez, M Gwinn and M.J Khoury, "Application of support vector machine modeling for prediction of common diseases: thecase of diabetes and pre-diabetes," BMC Medical Informatics and Decision Making, vol. 10, 2010.

[12] A. Liaw, M. Wiener et al., "Classification and regression by randomforest," R news, vol. 2, no. 3, pp. 18–22, 2002.

[13] Kumar, D. Vijaya, and VV Jaya Rama Krishniah. "An automatedframework forstroke and hemorrhage detection using decision treeclassifier." Communication and Electronics Systems (ICCES),International Conference on. IEEE, 2016.

[14] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," 2008 IEEE/ACS International Conference on Computer Systems and Applications, Doha, 2008, pp. 108-115. doi: 10.1109/AICCSA.2008.4493524.

[15] N. Liu, Z. Lin, Z. X. Koh, G.-B. Huang, W. Ser, M. E. H. Ong, "Patient outcome prediction with heart rate variability and vital signs", *J. Signal Process. Syst.*, vol. 64, pp. 265-278, 2011.

[16] A. Temko, C. Nadeu, W. Marnane, G. B. Boylan, G. Lightbody, "EEG signal description with spectral-envelope-based speech recognition features for detection of neonatal seizures", *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 6, pp. 839-847, Nov. 2011.

[17] S. J. Pan, Q. Yang, "A survey on transfer learning", *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.