Prediction of Heart Disease using Machine Learning Algorithms: A Survey

Himanshu Sharma,
Department of Computer Engineering and Applications,
National Institute of Technical Teachers' Training and Research.
hs13867@gmail.com

M A Rizvi,
Department of Computer Engineering and Applications,
National Institute of Technical Teachers' Training and Research.
marizvi@nitttrbpl.ac.in

***Abstract*:** According to recent survey by WHO organisation 17.5 million people dead each year. It will increase to 75 million in the year 2030[1].Medical professionals working in the field of heart disease have their own limitation, they can predict chance of heart attack up to 67% accuracy[2], with the current epidemic scenario doctors need a support system for more accurate prediction of heart disease. Machine learning algorithm and deep learning opens new door opportunities for precise predication of heart attack. Paper provideslot information about state of art methods in Machine learning and deep learning. An analytical comparison has been provided to help new researches' working in this field.

Keywords: *Machine learning, Heart Disease, Naïve Bayes, Decision Tree, Neural Network, SVM and Deep Learning.*

*****

I. INTRODUCTION

Heart disease has created a lot of serious concerned among researches; one of the major challenges in heart disease is correct detection and finding presence of it inside a human. Early techniques have not been so much efficient in finding it even medical professor are not so much efficient enough in predicating the heart disease[3]. There are various medical instrumentsavailable in the market for predicting heart disease there are two major problems in them, the first one is that they are very much expensive and second one is that they are not efficiently able to calculate the chance of heart disease in human. According to latest survey conducted by WHO, the medical professional able to correctly predicted only 67% of heart disease [2]so there is a vast scope of research in area of predicating heart disease in human.

With advancement in computer science has brought vast opportunities in different areas, medical science is one of the fields where the instrument of computer science can be used. In application areas of computer science varies from metrology to ocean engineering. Medical science also used some of the major available tools in computer science; in last decade artificial intelligence has gained its moment because of advancement in computation power. MachineLearning is one such tool which is widely utilized in different domains because it doesn't require different algorithm for different dataset. Reprogrammable capacities of machine learning bring a lot of strength and opens new doors of opportunities for area like medical science.

In medical science heart disease is one of the major challenges; because a lot of parameters and technicality is involve for accurately predicating this disease. Machine learning could be a better choice for achieving high accuracy for predicating not only heart disease but also another diseases because this vary tool utilizes feature vector and its various data types under various condition for predicating the heart disease, algorithms such as Naive Bayes, Decision Tree, KNN, Neural Network, are used to predicate risk of heart diseases each algorithm has its speciality such as Naive Bayes used probability for predicating heart disease, whereas decision tree is used to provide classified report for the heart disease, whereas the Neural Network provides opportunities to minimise the error in predication of heart disease. All these techniques are using old patient record for getting predication about new patient. This predication system for heart disease helps doctors to predict heart disease in the early stage of disease resulting in saving millions of life.

This survey paper is dedicated for wide scope survey in the field of machine learning technique in heart disease. Later part of this survey paper will discusses about various machine learning algorithm for heart disease and their relative comparison on the various parameter.It also shows future prospectus of machine learning algorithm in heart disease. This paper also does a deep analysis on utilization of deep learning in field of predicting heart disease.

II. LITERATURE REVIEW

Different researchers have contributed for the development of this field. Predication of heart disease based on machine learning algorithm is always curious case for researchers recently there is a wave of papers and research material on this area. Our goal in this chapter is to bring out all state of art work by different authors and researchers.

Marjia Sultana, Afrin Haider and Mohammad ShorifUddin[4] have illustrated about how the datasets available for heart disease are generally a raw in nature which is highly redundant and inconsistent. There is a need of pre-processing of these data sets; in this phase high dimensional data set is reduced to low data set. They also

show that extraction of crucial features from the data set because there is every kind of features. Selection of important features reduces work of training the algorithm and hence resulted in reduction in time complexity. Time is not only single parameter for comparison other parameters like accuracy also play vital role in proving effectiveness of algorithm similar. An approach proposed in[4] have worked to improved the accuracy and found that performance of Bayes Net and SMO classifiers are much optimal than MLP, J48 and KStar. Performance is measured by running algorithms (Bayes Net and SMO) on data set collected from WEKA software and then compared using predictive accuracy, ROC curve, ROC value.

Different methods have their own merits and demerits in work done by M.A.Jabbar, B.L Deekshatulu, Priti Chndra [5], an optimisation of feature has been done to achieve higher classification efficiency in Decision Tree . It is an approach for early detection of heart disease by utilizing variety of feature. These kind of approach can also be utilize for other sphere of research. Other than decision tree various other approach where adopt for achieving the goal of perfect detection of heart disease in human Yogeswaran Mohan et.al [6] have collected raw data form EEG device and used to train neural network for pattern classification . Here input output are depressive and non depressive categories in the hidden layer scaled conjugate gradient algorithm is used for training to achieve efficient result. authors have got efficiency up to 95% with help of trained neural network watching the success of neural network researches working in the domain of SVM have used this technique to classify and achieve more better result in case where the feature vector are multi dimensional and non linear these method defeated all other existing quantum contemporary techniques because it has capability to work under dataset of high dimensionality.

After going through majority of state of art technique we have pointed out certain loop holes existed in them. Some of them are discussed below

- There is wide need for more robust algorithm which can minimised the noise in the dataset because medical dataset may consists of various types of redundancy and noise in them.
- Recently with advancement in deep learning there could be chance to enhance efficiency and accuracy for detection heart disease
- Dimensionality of medical dataset is very high these put ergs to find such algorithm which can compress and reduce higher dimensionality, resulting in gaining execution time.

III. MACHINE LEARNING ALGORITHM FOR HEART DISEASE PREDICATION

Machine learning is widely used artificial intelligence tool in all major sector of application, with advancement in processing power machine for learning.

DECISION TREE

Decision tree is a graphical representation of specific decision situation that used for predictive model, main component of decision tree involves root, nodes, and branching decision. There are few approaches for building tree such as ID3, CART, CYT, C5.0 and J48 [7] has used the approaches to classify the dataset using J48, similarly [8] have compared decision tree with classification output of other algorithm. Decision tree is used in those area of the medical science where numerous parameters involved in classification of data set.

Since decision tree is most compressive approach among all machine learning algorithm. These clearly reflect important features in the data set. In heart disease where number of parameter affect patient such as blood pressure, blood sugar, age, sex, genetic and other factor. By seeing decision tree, doctor can clearly identifies the most effecting feature among all the parameter. They can also generate the most affecting feature in the mass of population. Decision tree is based on entropy and Information gain clearly signifies the importance of dataset. Drawback of decision tree is that it suffers from two major problems over fitting and it is based on greedy method. over fitting happened due to decision tree spilt dataset aligned to axis it means it need a lot of nodes to spilt data, this problem is resolved by J48 explained in[7]based on greedy method lead to less optimal tree, if dynamic approach is taken it lead to exponential number of tree which is not feasible.

SUPPORT VECTOR MACHINE (SVM)

A SVM performs classification by finding the hyper plane that maximise the margin between two classes. The vectors that define the hyper plane are the support vectors[9]

Steps for Calculation of Hyperplane

1. **Set up training data**
2. **Set up SVM parameter**
3. **Train the SVM**
4. **Region classified by the SVM**
5. **Support vector**

Usage of the SVM for data set classification has its own advantages and disadvantages. Medical data set can be non linear of high dimensionality by observing properties. It is clear that SVM would be one of the favourite choices for classification. Some of the advantage to select the SVM for classification choice.

1. Firstly regularisation parameters which avoid problem of over fitting which one of the major challenges is in decision tree.
2. Kernel tree is used to avoid the expert knowledge through the knowledge of kernel

3. SVM is an efficient method because it utilize convex optimisation problem (COP) which mean it has doesn't local minima
4. Error rated is tested which provide a greater support after misclassification of dataset

All the above features could be useful for medical diagnose dataset which is resulting in building more efficient predication system for the doctor. It doesn't mean it has all good side .coin has always two side on the other side it has best feature removal of over fitting problem is quite sensitive and it need optimizing parameter flaw in optimisation may result in error and may cause over fitting.

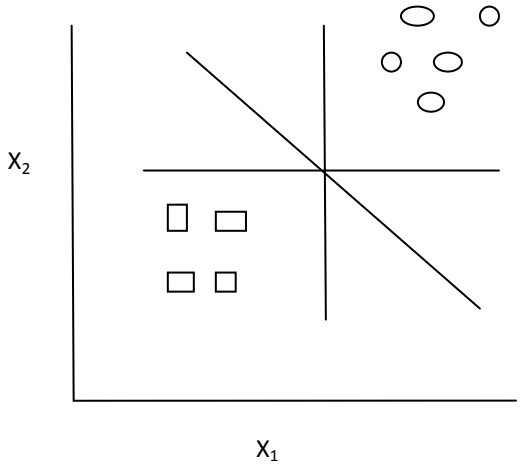

Fig 1 SVM Classifier

K- NEAREST NEIGHBOUR ALGORITHM (KNN)

KNN is slow supervised learning algorithm, it take more time to get trained classification like other algorithm is divided into two step training from data and testing it on new instance . The K Nearest Neighbour working principle is based on assignment of weight to the each data point which is called as neighbour. in K Nearest Neighbour distance is calculate for training dataset for each of the K Nearest data points now classification is done on basis of majority of votes there are three types of distances need to be measured in KNN Euclidian, Manhattan, Minkowski distance in which Euclidian will be consider most one the following formula is used to calculate their distance[10]:

$$Eucledian\ Distance = D(x,y) = (x_i - y_i)_{2k_i} = 1 \quad (1)$$

K=number of cluster

x , y=co-ordinate sample spaces

$$Manhattan\ distance = (x_i - y_i)_{n_{i=1}} \quad (2)$$

x&y are co-ordinates

Minkowski distances are generally Euclidian distance

$$Min = (|x_i - y_i|_p)_{1/p} \quad (3)$$

Grouping of sample is based on super class in the KNN reduction of sample is the result of proper grouping which is used for further training. Selection of k value plays a pivotal role, if the k value is large then it precise and less noisy.

The algorithm for KNN is defined in the steps given below

Rr

1. D represents the samples used in the training and k denotes the number of nearest neighbour.
2. Create super class for each sample class.
3. Compute Euclidian distance for every training sample
4. Based on majority of class in neighbour, classify the sample

IV. DEEP LEARNING FOR PREDICATION IN HEART DISEASE

Deep learning can be defined as subfield of machine learning which is based on learning at multiple level of representation and abstraction ,each level contains multiple processing unit for multiple processing between the input and output layer [10]. Deep learning work on the principle of feature hierarchy where higher level hierarchy is formed by composition lower level features. Deep learning bring renaissance to the neural network model major work is going in the field of in its implementation through stacked restricted Boltzmann machine and auto encoder-decoder technique[11]. This method impress researches with their performance in field of image processing and layer wise pre training techniques other areas of its application include Natural language processing,acoustic processing., RNN is consider to be best suited for sequential feature and sequential data their exist various method working on these two version LSTM was proposed by Hochreiter and Schmidhuber [12], the performance is quite impressive in the field related to sequence based task . Other contemporary method to LSTM is gated recurrent unit (GRU), it is simpler than LSTM but the result is quite impressive. A temporal based heart disease prediction has been done in paper [13]where author used GRE to achieve the high accuracy. Researchers have begun to use deep learning technique for medical dataset. Lasko et al. [14] is used encoder-decoder type pattern form serum of uric acid .similar kind of works have been discussed in great detail by the author. In generalised approach of deep learning has been illustrated in flow chart Fig. 2.

In the flow chart there are five modules it has their own specific operation, the goal is to present the above flow chart in most general way. Data collection is phase in which dataset from standard repository is get collected followed stage of pre-processing which include functionality of noise reduction and feature selection. Next step is core for deep learning because it implement the basic algorithmic approach adapted for manipulation of data set , the algorithms may vey from deep belief network[15] to

recurrent neural network . performance analysis of above data mining technique has been the major module because it illustrated about basic comparison of above adapted method , in the last discovery of knowledge module will get our desire goal which include percentage or probability of happening the instances. In our case it is the probability of heart attack in the patient

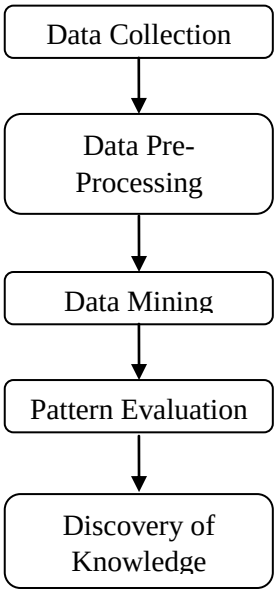

Fig 2 Flowchart of Deep Learning

V. ANALYSIS OF AVAILABLE LEARNING ALGORITHM

When it come to comparing two or more machine learning algorithm, it is most difficult because two algorithms is differ in many ways. Reason for difficulty in comparison because algorithm are highly depended on dataset , it is not wise to decide properly which algorithm is perform for the particular dataset , there is only one way to know about the efficiency of algorithm for the particular dataset is implement them. Analytical comparison is require to properly decide the difference between different machine learning algorithm this type ofwork could be useful for researchers who want to work in this fieldComparison will highlight the key difference on different background this paper has tried to reflect majority of comparison between different algorithms so that beginner and new .

Table 1. Compares major Machine learning Algorithm based on different parameter

Techniques	Outlier	Online learning	Over fitting and under fitting	Parametric	Accuracy	Execution Technique
SVM	It can handle outlier properly	Online training require less time than ANN	Perform better than over fitting and under fitting	Non parametric model	Higher than other parametric model	Depend upon dataset used, generally quite slow NLP operation
Decision Tree	Outlier does not play critical role in interoperation of dataset by decision tree	It does not supported online learning	It suffer over fitting and under fitting	Non parametric model	Accuracy depend on the dataset, ensemble technique used decision tree have higher accuracy than SVM	Require less time than other parametric model if not suffering from over fitting where as ensemble technique need higher execution than decision tree
Naïve Bays	It is less pruned to outlier	It can perform on online testing	It does not suffer over fitting and under fitting	It is parametric	High with limited dataset	Low with limited dataset
ANN	It is pruned to outlier	Online learning can take in ANN but more time than SVM	It is more pruned to over fitting than SVM	It is parametric	Higher than all other parametric model	Execution time depend upon number of layer declared and number of epochs need for testing
Linear Regression	It is less pruned to outlier because it strong problastic background	Require explicit training of classifier for new dataset	It does not suffer from under fitting and over fitting	It is parametric	Higher for linear dataset	Require less execution time than other model

researcher could get benefit.

Starting with Naive Bayes classifier it is quite easy to train classifier on small dataset if there exist high biasness and low variance give it major advantage over the classifier with low biasness and high variance such as KNN because later classifier will suffer problem of over fitting. The training on small dataset is due to the reason it converse very quickly so need less training data as well as less training time but as we all know that every coin has two side if data size started growing there is chance of asymptotic error where as the algorithm with low biasness and low variance are powerful enough to avoid this kind of problem .the other major disadvantage of Naïve Bayes algorithm is that it cannot learn interaction between features. On the other hand if considering the logistic regression model take care of related feature unlike the Naïve Bayes. Logistic Regression will also provide a firm mathematical probabilistic approach but

if data type is non linear the logistic regression model fail to provide any output. Hence it requires lot of feature modulation before feeding the dataset to model which quite teasing. But it is quite user friendly to update the mode if the feature in the dataset of linear type even if new rows and column arrives with the time. i.e. that it is perform quite well with online dataset and temporal dataset .

Decision tree is non parametric machine learning algorithm which is considerable if the compressibility is the major feature because it is quite easy to explain the model internal and external architecture. Decision tree suffers from some of the major drawback such as it does not support online learning and suffers from over fitting of dataset but there exist quite few technique which can avoid the over fitting such as J48 model. Ensemble technique such as random forest[16] can provide quite a few impugned in decision tree such as it solve the problem of imbalance dataset, pruning and an accuracy it is said that random forest has potential to replace most accurate mode of machine learning algorithm but it is snatch the properties of compressibility of the decision tree.

SVM and Neural network are consider to be major competitive machine learning algorithm but they are very much difference from each other with the same goal of classifications or regression both of them are non linear classification technique.

SVM is derived from algebraic and statics background it construct a linear separable hyper plane in n dimensional space to separate all classifier with large margin it is consider theoretical that SVM will provide high accuracy to each dataset with high dimensionality. ANN is also non linear model it suffer a lot of drawbacks which SVM avoids such as SVM converge only on global and unique minima on the other hand ANN converge on each local minima, since SVM has quite good mathematical background it can be represented geometrical that does not exist any such geometrical representation of ANN model, it is also point to be noted ANN complexity depend a lot of dimensionality of dataset where as SVM is free of such problem.

It does not mean SVM can overshadow every other algorithm it has its own limitation, SVM is very hard to interrupt and tune because it is memory intensive, SVM is not easily for training of NLP based method because hundred of thousand feature get created in these which will result exponentially increase in time complexity where as ANN model still give linear result. ANN also outperforms SVM for online training of dataset. certain parameter along with difference model have been relatively compared in the tabular format in given below table which reflect the drawback and advantages of every algorithm on each parameters

VI. CONCLUSION

Heart attack is crucial health problem in human society. This paper has summarised state of art techniques and available methods for predication of this disease. Deep learning an emerging area of artificial intelligence showed some promising result in other field of medical diagnose with high accuracy. It is still an open domain waiting to get implemented in heart diseasepredication.Some methods of deep learning has been discussed which can be implemented for heart disease predication, along with pioneer machine learning algorithms. An analytical comparison has been done for finding out bestavailable algorithm for medical dataset. In future our aim is to carry forward the work of temporal medical dataset, where dataset varies with time and retraining of dataset is required

REFERENCES

[1] William Carroll; G. Edward Miller, “Disease among Elderly Americans : Estimates for the US civilian non institutionalized population, 2010,” *Med. Expend. Panel Surv.*, no. June, pp. 1–8, 2013.

[2] V. Kirubha and S. M. Priya, “Survey on Data Mining Algorithms in Disease Prediction,” vol. 38, no. 3, pp. 124–128, 2016.

[3] M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, “Prediction of risk score for heart disease using associative classification and hybrid feature subset selection,” *Int. Conf. Intell. Syst. Des. Appl. ISDA*, pp. 628–634, 2012.

[4] M. Sultana, A. Haider, and M. S. Uddin, “Analysis of data mining techniques for heart disease prediction,” *2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. iCEEiCT 2016*, 2017.

[5] M. Akhil, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,” *Procedia Technol.*, vol. 10, pp. 85–94, 2013.

[6] S. Kumra, R. Saxena, and S. Mehta, “An Extensive Review on Swarm Robotics,” pp. 140–145, 2009.

[7] T. M. Lakshmi, A. Martin, R. M. Begum, and V. P. Venkatesan, “An Analysis on Performance of Decision Tree Algorithms using Student’s Qualitative Data,” *Int. J. Mod. Educ. Comput. Sci.*, vol. 5, no. 5, pp. 18–27, 2013.

[8] P. Sharma and A. P. R. Bhartiya, “Implementation of Decision Tree Algorithm to Analysis the Performance,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 1, no. 10, pp. 861–864, 2012.

[9] D. K. Srivastava and L. Bhambhu, “Data classification using support vector machine,” *J. Theor. Appl. Inf. Technol.*, 2009.

[10] N. Bhatia and C. Author, “Survey of Nearest Neighbor Techniques,” *IJCSIS) Int. J. Comput. Sci. Inf. Secur.*, vol. 8, no. 2, pp. 302–305, 2010.

[11] J. Schmidhuber, “Deep Learning in neural networks: An overview,” 2015.

[12] S. Hochreiter and J. Urgen Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," *2008 IEEE/ACS Int. Conf. Comput. Syst. Appl.*, pp. 108–115, 2008.

[14] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data," *PLoS One*, vol. 8, no. 6, 2013.

[15] Yuming Hua, Junhai Guo, and Hua Zhao, "Deep Belief Networks and deep learning," *Proc. 2015 Int. Conf. Intell. Comput. Internet Things*, pp. 1–4, 2015.

[16] P. De, "Modified Random Forest Approach for Resource Allocation in 5G Network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 11, pp. 405–413, 2016.