Review of Heart Disease Prediction using Supervise and Unsupervised Machine Learning Technique

[1]Pooja Gupta, [2]Pritesh Jain, [3]Upendra Singh

[1]M.Tech Scholar, [2]Assistant Professor
[1,2]Patel College of Science and Technology, Indore

***Abstract*: As per the records of WHO 17.5 million people die every year. By the year 2030 it will rise up to 75 million [1]. The medical professionals associated with heart diseases have some limitations, they can predict the chances of heart attack with the accuracy of 67% [2], for more accurate predictions of heart diseases, and doctors require a support system. The precision in predictions of heart attack can be achieved by deep and machine learning algorithms. In this paper there is a lot of information about the state of art methods in deep learning and machine learning. To assist new researcher's active in this area an analytical comparison has been provided.**

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

I. INTRODUCTION

The diseases of heart have created a lot of grave distress among researchers; one of the chief concerns in heart diseases is the accurate detection and the presence of this in a human being. The techniques of early age have not been much competent in detecting it, even medical professors are less efficient in predicting the heart diseases [3]. There a number of medical instruments available in the market but there are two main drawbacks; a). the instruments are very high-priced, and b). They are not efficient enough to calculate the heart diseases. As per the figures of latest survey by WHO, the medical professionals are only capable to predict 67% of heart diseases [2]. Therefore, scope and scale of research in this area is very high and large respectively.

The field of computer science has advanced at an incredible rate and has opened a huge number of opportunities in different areas of science and technology. The medical science is one such areas where the instruments of computer science can be utilized. The applied part of computer science varies from ocean engineering to meteorology.

Some of the main existing tools in computer science has been used by various medical sciences, for example, due to the rapid advancement in computation power the artificial intelligence is now reached the zenith of its existence. Machine learning is one of the tools that is wholly available as it does not require different algorithm for various datasets. The reprogrammable capacities of machine learning bring a lot of opportunities for medical sciences.

For accurately predicating the heart diseases a great number of parameters and complex technicality is involved. This is a challenge faced by the medical sciences. Machine Learning can play a vital role in facing such challenges because its accuracy is impeccable. For predicating heart diseases, this method of learning uses varied tools such as feature vector and its different data types under various conditions.

The risk of heart diseases can be predicted by the algorithms such as Decision Tree, Naive Bayes, Neural Network and KNN. Every algorithm has its speciality, for example, Naive Bayes used probability to predict heart diseases whereas Decision Tree provides classified reports for the same, and the Neural Network lessens the margin of error in predication of heart diseases. The old records of heart-patients are being used by these techniques to get accurate predictions of new patients. These predictions help doctors to save millions of lives.

This paper is dedicated to provide information about the scope of machine learning technique in heart diseases. As we go further, this paper discusses about different machine learning algorithm and their comparative parameters. It also illustrates upcoming prospects of machine learning algorithm and its deep analysis.

II. LITERATURE REVIEW

Researchers from various scientific backgrounds have contributed to develop this field. This machine-learning based prediction has always been one of the most curious research areas for science fraternity. There is a sudden rise in researchers working on the papers and materials associated with this area. Our chief goal is to provide all the state of artworks by various authors and researchers. Afrin Haider, Mohammad Shorif Uddin and Marjia Sultana [4] have illustrated that the datasets for heart diseases are raw and highly superfluous and incoherent in nature. There is an urgent need of pre-processing of datasets because in this phase the high-dimensional dataset is reduced to low dataset.

There are all sorts of features available in a dataset and they show the extraction of some crucial features. Reduction in the work of training the algorithm depends mainly upon a vital factor, i.e., selection of significant features. It results in reducing time complexity. To prove the effectiveness of algorithm the two vital parameters for comparison are used; a). Time, b). Accuracy. An effective approach has been proposed in[4] and it has contributed to ameliorate the accuracy, precision and has discovered a fact that performances of Bayes Net and SMO classifiers are much optimal compared to J48, KStar and MLP. The performance is calculated by running algorithms (Bayes Net and SMO) on data set which was collected from a WEKA software and then put into comparison by using ROC curve, ROC value and predictive accuracy. Various methods have their own merits and demerits in work-done by M.A. Jabbar, Preeti Chandra and B.L Deekshatulu [5]. In order to achieve efficiency of higher class in a Decision Tree the optimisation of feature has been performed. By utilizing various features early detection of heart disease can be done. These sorts of approaches can also be utilized in other spheres of research. Some approaches other than decision tree which are dedicated to achieve the goal of perfect detection of heart disease in humans is explained by Yogeswaran Mohan et.al [6] who have collected raw data form EEG devices and made use of it in training neural-network for pattern classifications. Input and output are the depressive and non-depressive characteristics of the hidden layer in which a scaled conjugate gradient algorithm is used for training and achieving efficient results.

Some authors have achieved 95% efficiency with the help of trained neural networks. Researchers who are working in the field of SVM have watched the success of neural network and used this technique to classify and achieve more advanced and enhanced results. The Neural Network has the capability to work under high dimensional dataset. When the feature vector which are multi-dimensional and non-linear come into play this method defeats all other existing quantum contemporary techniques.

We have pointed out certain loop holes after going through majority of state of art techniques. Some of them are as follows:
a) Due to various types of redundancy and noise in medical dataset there is a huge demand for more robust algorithms which can reduce the noise.
b) There is a chance of enhancement in the efficiency and accuracy of detection of heart diseases with the help of recent advancement in the field of deep learning.
c) Due to very dimensionality of medical dataset there are ergs to find such algorithms which can reduce and compress higher dimensionality results in gaining more execution time.

III. MACHINE LEARNING ALGORITHM FOR HEART DISEASE PREDICATION

With advancement in processing power Machine Learning is a tool of artificial intelligence which is widely used in all the chief segments of application. Decision Tree is a graphical illustration of an exact decision that is used for model of predication. Nodes, roots and branching decisions are the main components of a Decision Tree. CART, ID3, CYT, C5.0 and J48 [7] are some approaches to build a tree. These have used the approaches to categorize the dataset by using J48, similarly [8] decision tree have been compared with the classification output of various algorithm. In medical sciences, when numerous parameters are involved in classifying a data set Decision Tree comes into play.
As the most compressive approach among all machines learning algorithm, Decision Tree reflects important features in the data. There are many parameters which affect the patient in a heart disease such as blood sugar, blood pressure age, sex, genetic and other factor. The decision tree assists doctors to evidently identify the feature which affects the most. Among the mass of population they can also easily generate the most affecting feature. The importance of dataset can be witnessed in the Decision Tree which is completely based upon entropies and information. Over fitting and greedy methods are the two main drawbacks of Decision Tree. The reason that caused over fitting was the decision tree spilt dataset aligned to axis, we can understand that it needs a lot of nodes to spilt data. Based on greedy methods which leads to less optimal tree, this problem is resolved by J48 explained in[7]. If the dynamic approach is taken into consideration it might lead to the exponential number of tree which is not reasonable.

SUPPORT VECTOR MACHINE (SVM)

By finding the hyper plane which maximises the margin between two classes the classification of SVM is performed. Support vectors [9] are the ones that define those hyper-planes. The Steps for Calculation of Hyper-plane are as follows:

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

There are various advantages and disadvantages of the usage of SVM dataset classification**.** By observing properties a medical dataset can be non-linear of high dimensionality. It is one of the most widely accepted facts that SVM is the great choice for classification. Some of the advantages are :

1. At first, the regularisation of parameters which keep away from problems of over-fitting which, generally, is one of the major challenges in a decision tree.

2. Basically, a Kernel tree is used to ignore the expert knowledge through the kernel knowledge.

3. The SVM is a proficient process as it utilizes convex-optimisation-problem (COP), which means it
Does not have local minima.

4. When misclassification of dataset happens an Error Rate is put into testing which is a sort of great support. These features are useful in medical diagnosis which, ultimately, builds more proficient predication system. It also does not mean that it has all the goods in it. A coin does always have two sides. On the other side it has Some very good features that eliminates the over fitting problem which is quite sensitive and needs to get an optimized parameter flaw.
Sometimes, optimisation might result in errors and may cause over fitting.

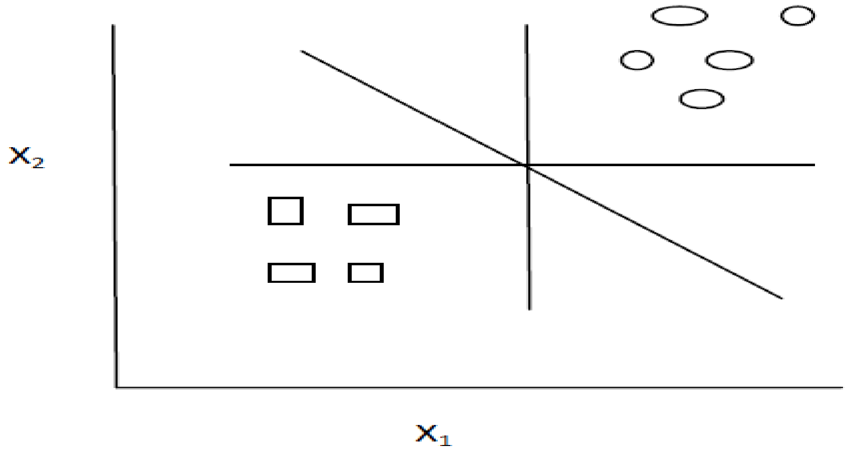

Fig 1 SVM Classifier

K- NEAREST NEIGHBOUR ALGORITHM (KNN)

KNN is a learning algorithm which is very slow and supervised. Comparatively, it takes more time to achieve trained classifications. Like the other algorithms are divided into two steps; a). Training from data and, b). Testing it on new instances. The working principle of K-Nearest Neighbour is based upon an assignment of weight to each point of data which is known as neighbour. In this classifier, the distance is calculated for training the dataset for every K Nearest data points. Now, it is classified on the basis of widely held number of votes. Three types of distances that are required to be measured in KNN are Minkowski, Euclidian and Manhattan.

The formula below is used to calculate their distances [10]:
$Eucledian\ Distance = D\ x, y = (xi - yi)2ki = 1$ (1)
K=number of cluster
x , y=co-ordinate sample spaces
$Manhattan\ distance = (xi - \ )$=1 (2)
x&y are co-ordinates
Minkowski distances are generally Euclidian distance $Min = (\ - yi\ p\ )1\ p$ (3)
The sample grouping is based upon the super class in the KNN.
The result of proper grouping is the reduction of sample which is further utilized for training. Selection of the value of k plays an essential role, if the k value is large, then it is less noisy and precise. The KNN algorithm is defined in the following steps:

1. k denotes the number of nearest neighbour and D represents the samples utilized in the training.
2. For each sample class create a super class.
3. For every training sample compute Euclidian distance
4. Classify the sample based on majority of class in the neighbour.

IV. DEEP LEARNING FOR PREDICATION IN HEART DISEASE

Based upon learning at multiple level of abstraction and representation deep learning can be defined as subfield of machine learning, there are multiple processing unit between an input and output layer [10]. Deep learning is such an algorithm which works on the principles of feature-hierarchy. Here, the composition of lower level features forms the higher level hierarchy.
The deep learning brought the renaissance to the neural network models. There are lot many major work which is going on in this field by implementing stacked restricted Boltzmann Machine and auto encoder-decoder technique [11].
The researchers are impressed by the performance of this method. Performance of image processing layer wise pre-training techniques were also the areas of interest. Natural language processing and acoustic processing are the other areas of interest. For sequential feature and data, RNN is considered to be the best. Various methods are applied for the above two versions, some are

LSTM which was proposed by Hochreiter and Schmidhuber [12]. In the sequence based task their performance is very much appreciated.

Gated Recurrent Unit (GRU) is the other modern technique of LSTM. The results of GRU are quite impressive and it's simpler that LSTM.

There is a paper [13] in which a sequential heart disease prediction has been discussed. To achieve high accuracy authors have utilized GRE. The deep learning technique is being used for medical dataset by the modern day researchers. From the serum of uric acid Lasko et al. . [14] Utilized encoder-decoder pattern. The illustration of generalised approach of deep learning is in the flow chart of Fig. 2.

There are five types of modules present in a flow chart. Every module has a specific operation. The collection of dataset from the standard repository is called Data Collection. It is then followed by a pre-processing stage where functionality in reduction of noise, and feature selection are included. The core for deep-learning is the next step because the implementation of the vital algorithmic approach adapted for manoeuvring of dataset is present. The algorithms might vary from recurrent neural network to deep belief network [15].

The data mining technique above went through an analysis of performance which became the main module as it has depicted the basic comparison of the methods explained above.

The modules, at the final discovery of virtue, will get expected results, like, the probability or percentage of the instances taking place. Here, in this scenario, it is the probability of heart attacks taking place in different types of patients.

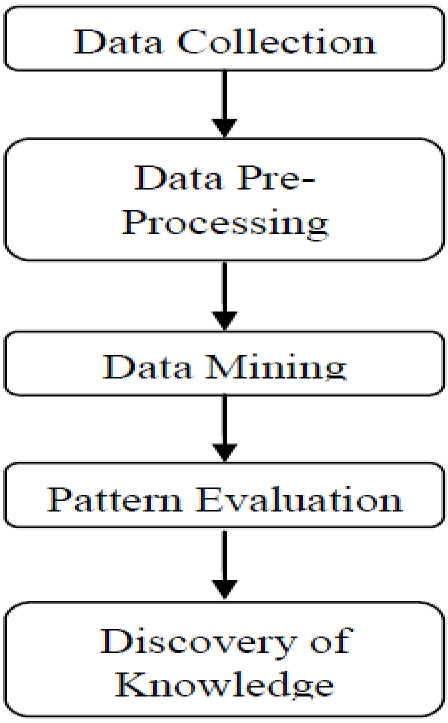

Fig 2 Flowchart of Deep Learning

V. ANALYSIS OF AVAILABLE LEARNING ALGORITHM

Two algorithms differ from each other in innumerable ways. So it becomes difficult to compare two or more machine learning algorithms. The reason behind this is the complete dependency of algorithms on dataset. This results in complicating the decision making strategy, when the performance of an algorithm for an individual dataset is talked about. Only by implementing the algorithms with a particular dataset we can find out the efficiency of algorithms. The analytical comparison is required to take an appropriate decision in order to differentiate between various machine-learning algorithms. These types of works can be helpful for the researchers who want to work in this field of comparisons. This paper has made an effort to reflect majority of the comparisons between various algorithms, so that beginners and new researchers could get some advantage.

Techniques	**Outlier**	**Online learning**	**Over fitting and under fitting**	**Parametric**	**Accuracy**	**Execution on 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 support ed 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
Naive 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 prune d 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

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

Naive Bayes classifier explains that when there exists a high-biasness and low-variance the training of classifier on small dataset becomes effortless and is full of advantages in comparison with the classifier that has low-biasness and high-variance, such as KNN. It is because the classifier, later, suffers the problem of over fitting. The training on small dataset is due to the reason it converts very quickly to needless data and time. But as we are aware of the fact that every coin has two sides, if the size of the data starts growing, there are chances of asymptomatic errors whereas the algorithm that has low biasness and low variance are powerful enough to avoid these types of problems.

There are some other main disadvantages of Naive Bayes algorithm, for example, it cannot learn interaction among various features. In contrary, if the logistic regression model considers taking care of associated feature unlike the Naive Bayes, Logistic Regression tends to provide a certain mathematical probabilistic approach. But in cases where the data type is non-linear the logistic regression model fails to provide any output. Therefore, before feeding the dataset to model it requires a lot of feature modulation which can be quite teasing. But it has always been friendly to the users to update the mode of the feature in the dataset of linear type, even though, the new rows and column arrives with time, i.e., it executes well with temporal and online dataset. The Internal and external architecture of the models can be easily explained if the substantial compressibility is the major feature in the Decision Tree which is a non-parametric machine learning algorithm. There are some unfortunate drawbacks of the decision tree, like, online learning is not supported and it suffers from the over-fitting of the dataset. But there are some techniques such as J48 model which avoids over-fitting. Random forest is a an ensemble technique[16] which provides a few impugned in a decision tree, e.g., accuracy, pruning and solves the problems of an imbalanced dataset. The random forest is believed to have the potential to replace most accurate modes of machine learning algorithms but there is a drawback, it seizes the compressible property of the decision tree. It is considered that that Neural network and SVM are the two main competitive machine learning algorithms. But they are, actually, very different from each other with the similar motive of classification or regression. These two are non-linear classification techniques. From the derivation of statics and algebra we get SVM which constructs linear separable hyper plane in N dimensional plane in order to separate all the classifiers which has large margins. It is theoretically considered that SVM provides high level of accuracy to each dataset with high dimensionality. The ANN is also one the non-linear models which has plenty of drawbacks among which one is that ANN converge on each local minima. Generally, SVM avoids such dilemmas and converge on global and unique minima. The SVM can represent geometrically as it comes from a fine mathematical background. The representation of ANN model is no match to the SVM model because the complexity of ANN depends a lot upon dimensionality of dataset whereas SVM is devoid of these problems.

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

The heart attack cases are increasing rapidly and have become a major concern in the human society. The state of art techniques and available methods for predication of this disease have been summarised in this paper. Deep learning and artificial intelligence has shown some incredible results in different areas of medical diagnosis with high accuracy. This domain is still waiting to get implemented in the predication of of heart disease. Along with pioneer machine learning algorithms some processes of deep learning have been considered which can be implemented for heart disease predication. For finding out the best available algorithm for medical datasets an analytical comparison has been done. In future, our main objective will be the ascension of the work of temporal medical dataset, where dataset varies in accordance with time and re-training of dataset is necessitated.

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, nd 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.

[17] Ashish Sharma ,Dinesh Bhuriya ,Upendra Singh, "Survey Of Stock Market Prediction Using Machine Learning Approach" , Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.

[18] Dinesh Bhuriya ,Girish Kaushal ,Ashish Sharma," Stock Market Predication Using A Linear Regression ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp. 1-4.

[19] Rohit Verma ,Pkumar Choure ,Upendra Singh , "Neural Networks Through Stock Market Data Prediction" , Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-6.

[20] Sonal Sable ,Ankita Porwal ,Upendra Singh , "Stock Price Prediction Using Genetic Algorithms And Evolution Strategies ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.

[21] Vineeta Prakaulya ,Roopesh Sharma ,Upendra Singh, "Railway Passenger Forecasting Using Time Series Decomposition Model ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.

[22] Yashika Mathur ,Pritesh Jain ,Upendra Singh, "Foremost Section Study And Kernel Support Vector Machine Through Brain Images Classifier ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-4

[23] Pooja Kewat , Roopesh Sharma , Upendra Singh , Ravikant Itare, "Support Vector Machines Through Financial Time Series Forecasting ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp. 1-7.