A Survey In Predicting Heart Disease At Early Stages Using Machine Learning

Evalin Jose[1] and Thanzeela Mol A[2]
PG Scholar[1] and Assistant Professor[2]
Al Ameen Engineering College, Shoranur, Kerala, India

Abstract: *Predicting and detection of heart condition has always been a critical and challenging task for healthcare practitioners. Hospitals and other clinics are offering expensive therapies and operations to treat heart diseases. So, predicting heart condition at the first stages are going to be useful to the people round the world in order that they're going to take necessary actions before getting severe. heart condition may be a significant problem in recent times; the most reason for this disease is that the intake of alcohol, tobacco, and lack of workout . Over the years, machine learning shows effective leads to making decisions and predictions from the broad set of knowledge produced by the health care industry. a number of the supervised machine learning techniques utilized in this prediction of heart condition are artificial neural network (ANN), decision tree (DT), random forest (RF), support vector machine (SVM), naïve Bayes (NB) and k nearest neighbor algorithm. Furthermore, the performances of those algorithms are summarized to seek out the higher accuracy obtained machine learning algorithm.*

I. Introduction

The heart is one among the most parts of the physical body after the brain. the first function of the guts is to pumping blood to the entire body parts. Any disorder which will cause disturbing the functionality of the guts is named heart condition . Several sorts of heart condition are there within the world; arteriacoronaria disease (CAD), and coronary failure (HF) are the foremost common heart diseases that are present. the most reason behind the CAD is blockage or narrowing down of the coronary arteries. Coronary arteries also are liable for supplying blood to the guts. CAD is that the leading explanation for death over 26 million people are affected by coronary heart condition CAD round the world, and it's increasing 2% annually thanks to CAD 17.5 million deaths happened globally in 2005.

Different factors can raise the danger of coronary failure . Medical scientists have classified those factors into two different categories; one among them is risk factors that can't be changed, and another one is risk factors which will be changed. case history, sex, age comes under risk factors that cannot be changed. High cholesterol, smoking, physical inactivity, high blood pleasure all come under risk factors.

Heart disease may be a significant issue, so there's a requirement for diagnosis or prediction of heart condition there are several methods to diagnose heart condition among them Angiography is that the trending method that's employed by most physicians across the planet . However, there are some drawbacks related to the angiography technique. it's an upscale procedure and physicians need to analyze numerous factors to diagnose a patient hence this process makes physician jobs very difficult, so these limitations motivate to develop a noninvasive method for the prediction of heart condition . These conventional methods affect medical reports of the patients moreover these conventional methods are time consuming, and that they may give erroneous results because these conventional methods are performed by humans. To avoid these errors and to realize better and faster results, we'd like an automatic system. Over the past years, researchers determine that machine learning algorithms perform alright in analyzing medical data sets.These data sets are going to be directly given to machine learning algorithms, and machine learning algorithms will perform consistent with their nature, and people algorithms will give some outputs.

II. Machine Learning Algorithms

Machine learning could also be a kind of AI that allows a system to seek out out from data rather than through explicit programming. However, machine learning isn't an easy process. because the algorithms ingest training data, it's

DOI: 10.48175/IJARSCT-1251

then possible to provide more precise models supported that data. A machine learning model is that the output generated once you train your Machine learning algorithm with data. After training, once you provide a model with an input, you'll tend an output.

Machine learning enables models to teach on data sets before being deployed. Some machine learning models are online and continuous. This iterative process of online models leads to an improvement within the kinds of associations made between data elements. Due to their complexity and size, these patterns and associations could have easily been overlooked by human observation. After a model has been trained, it are often utilized in real time to seek out out from data. The improvements in accuracy are a results of the training process and automation that are a neighborhood of machine learning.

Machine learning techniques are required to reinforce the accuracy of predictive models. counting on the character of the business problem being addressed, there are different approaches supported the type and volume of the data.

- **Supervised learning**: Supervised learning typically begins with a longtime set of data and a specific understanding of how that data is assessed. Supervised learning is meant to seek out patterns in data which will be applied to an analytics process. This data has labeled features that outline the meaning of data.
- **Unsupervised learning**: Unsupervised learning is used when the matter requires an enormous amount of unlabeled data. for instance, social media applications, like Twitter, Instagram, and Snapchat, all have large amounts of unlabeled data. Understanding the meaning behind this data requires algorithms that classify the data supported the patterns or clusters it finds. Unsupervised learning conducts an iterative process, analyzing data without human intervention. it's used with email spam detecting technology. There are far too many variables in legitimate and spam emails for an analyst to tag unsolicited bulk emails. Instead, machine learning classifiers, supported clustering and association, are applied to identify unwanted emails.
- **Reinforcement learning**: Reinforcement learning could also be a behavioral learning model. The algorithm receives feedback fromthe info analysis, guiding the user to the simplest outcome. Reinforcement learning differs from other kinds of supervised learningbecause the system isn't trained with the sample data set. Rather, the system learns through trial and error. Therefore, a sequence of successful decisions will end within the method being reinforced, because it best solves the matter at hand. Now we'll undergo different machine learning algorithms.

2.1 Artificial Neural Network

An artificial neural network (ANN) is that the piece of a computer system designed to simulate the way the human brain analyzes and processes information. It is the inspiration of AI (AI) and solves problems which may prove impossible or difficult by human or statistical standards. ANNs have self-learning capabilities that enable them to provide better results as more data becomes available.

Artificial neural networks are built a touch just like the human brain, with neuron nodes interconnected quite an online. The human brain has many billions of cells called neurons. Each neuron is formed from a cell body that's liable for processing information by carrying information towards (inputs) and away (outputs) from the brain.

An ANN has hundreds or thousands of artificial neurons called processing units, which are interconnected by nodes. These processing units are made from input and output units. The input units receive various forms and structures of data supported an indoor weighting system, and therefore the neural network attempts to find out about the knowledge presented to supply one output report. Just like humans need rules and guidelines to return up with a result or output.

ANN also uses a gaggle of learning rules called back propagation, an abbreviation for backward propagation of error, to perfect their output results. The Neural Network is made from 3 sorts of layers:

- **Input layer**: Initial data for the neural network.
- **Hidden layers**: Intermediate layer between input and output layer and place where all the computation is completed .
- **Output layer**: Produce the result for given inputs.

DOI: 10.48175/IJARSCT-1251

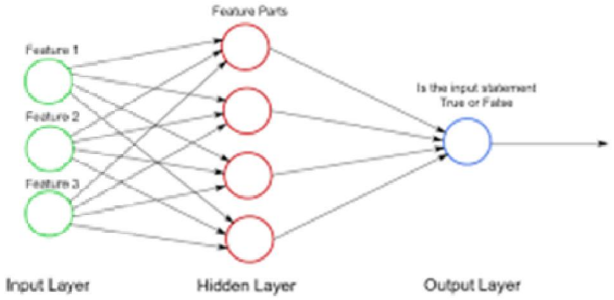

Figure 2.1: Artificial Neural Network

2.2 Support Vector Machine

Support Vector Machine (SVM) could also be a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as some extent in n dimensional space (where n is that the amount of features you have) with the price of every feature being the price of a selected coordinate. Then, we perform classification by finding the hyper plane that differentiates the 2 classes alright.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where the amount of dimensions is bigger than the amount of samples.
- Uses a subset of coaching points within the choice function (called support vectors), so it is also memory efficient.

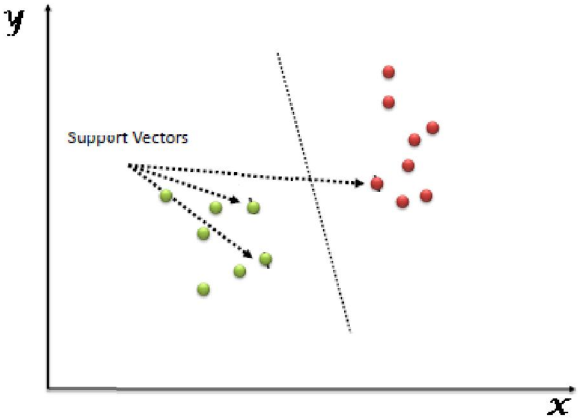

Figure 2.2: Support Vector Machine

- Versatile: different Kernel functions are often specified for the choice function. Common kernels are provided, but it's also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the quantity of features is way greater than the quantity of samples, avoid overfitting in choosing Kernel functions and regularization term is crucial.
- SVMs don't directly provide probability estimates, these are calculated using an upscale fivefold cross validation.

2.3 Decision Tree

Decision Trees are a kind of Supervised Machine Learning where the info is continuously split consistent with a particular parameter. The tree are often explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the outcomes. And the decision nodes are where the info is split. Decision trees are constructed via an algorithmic approach that identifies ways to separate a knowledge set supported different conditions. It is one among the foremost widely used and practical methods for supervised learning. Decision Trees are a non parametric supervised learning method used for both classification and regression tasks. Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

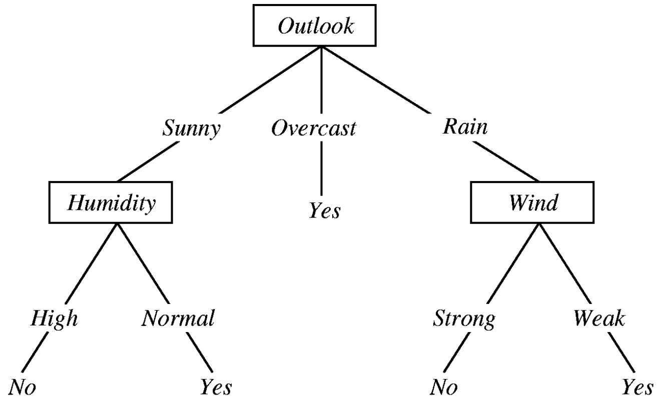

Figure 2.3: Decision Tree

2.4 Random Forest

Random forest could also be a versatile , easy to use machine learning algorithm that produces, even without hyper parameter tuning, a superb result most of the time. It is also one of the foremost used algorithms, thanks to its simplicity and variety. Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a mixture of learning models increases the general result.

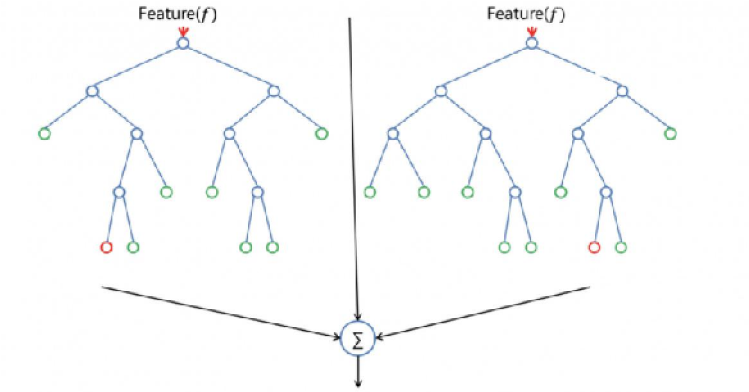

Figure 2.4: Random Forest

One big advantage of random forest is that it are often used for both classification and regression problems, which form the bulk of current machine learning systems. The Random forest has nearly an equivalent hyperparameters as a choice tree or a bagging classifier. Fortunately, there's no need to combine a choice tree with a bagging classifier because you'll easily use the classifier class of random forest. With random forest, you'll also affect regression tasks by using the algorithm's regressor. The Random forest has nearly an equivalent hyperparameters as a choice tree or a bagging classifier. Fortunately, there's no need to combine a choice tree with a bagging classifier because you'll easily use the classifier class of random forest. With random forest, you'll also affect regression tasks by using the algorithm's regressor.

Another great quality of the random forest algorithm is that it's very easy to live the relative importance of every feature on the prediction. Sklearn provides an excellent tool for this that measures a feature's importance by watching what proportion the tree nodes that use that feature reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results therefore the sum of all importance is capable one.

2.5 Naive Bayes

It is a classification technique supported Bayes' theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a specific feature during a class is unrelated to the presence of the other feature. For example, a fruit could also be considered to be an apple if it's red, round, and about 3 inches in diameter. Even if these features depend upon one another or upon the existence of the opposite features, all of those properties independently contribute to the probability that this fruit is an apple which is why it's referred to as ‘Naive’. The Naive Bayes model is straightforward to create and particularly useful for very large data sets. Along with simplicity, Naive Bayes is understood to outperform even highly sophisticated classification methods.

Applications of Naive Bayes Algorithms

- Real-time Prediction: Naive Bayes is an eager learning classifier and it is sure fast. Thus, it might be used for creating predictions in real time.
- Multiclass Prediction: This algorithm is also well known for its multiclass prediction feature. Here we will predict the probability of multiple classes of the target variable.
- Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers mostly used in text classification have a higher success rate as compared to other algorithms. As a result, it's widely utilized in Spam filtering and Sentiment Analysis

III. Literature Survey

LIAQAT ALI et al. recommended a model which is consists of two methods one is X^2 statistical and deep neural network (DNN). Feature refinement is done by X^2 statistical model and classification is done by a deep neural network(DNN). In their study, they have used the Cleveland dataset. There are 303 instances in that dataset, among them, 297 have no missing data, and the remaining 6 have missing data. Among 297, 207 instances are used for training data, and the remaining 90 are used as testing data. This model gives better results compared to conventional ANN models which are present earlier. As a result of using this proposed model, they have got 93.33% classification accuracy using DNN. It is 3.33% more than that of the conventional ANN model. The strength ofthe proposed diagnostic system was evaluated using six different evaluation metrics including accuracy, sensitivity, specificity, MCC,AUC, and ROC charts. Moreover, the performance of the proposed method was compared with other well known machine learning models and with other methods discussed in the literature. From the experimental results, we can safely conclude that the proposed diagnostic system can improve the quality of decision making during the diagnosis process of heart disease. The proposed method achieved higher detection accuracy for HF disease, but the current study did not investigate the time complexity of the proposed hybrid diagnostic system. In future studies, it will be investigated as it is considered an important factor in clinical application. Another limitation of the current study is that the optimal width of each hidden layer in the ANN and DNN model is searched using a grid search algorithm.

DR.KANAK SAXENA et al. developed a data mining model to predict heart disease efficiently. It mainly helps the medical practitioners to make efficient decisions way based on the given parameters. The author has used the Cleveland dataset from UCI, and they have used age, sex, resting blood pressure, chest pain, serum cholesterol, fasting blood sugar, etc. as attributes. Furthermore, they have divided the datasets into two parts one is for testing, and the other one is for training. They have used a 10fold method to find accuracy and find the accuracy of 86.3 % in the testing phase and 87.3 % in the training phase and because this model demonstrates better results and helps the area specialists and even individual related to the field to get ready for a superior determine and give the patient to have early determination results because it performs sensibly well even without retraining.

AWAIS NIMAT et al. proposed an expert system based on two support vector machines(SVM) to predict heart disease efficiently. These two SVM's have their purpose; first, one is used to remove the unnecessary features, and the second one is used for prediction. Moreover, they have used the HGSA (hybrid gird search algorithm) to optimize the two methods. By using this model, they have achieved 3.3% better accuracy than the conventional SVM models that are present earlier. proves the strength of the conventional SVM model by 3.3%. Moreover, the proposed method is capable of showing better results with a few features. Thus, the proposed method is efficient in terms of time complexity as well. Because it reduces the training time of the predictive model. Hence, from the experimental results achieved on the heart failure dataset, it is concluded that the proposed expert system can improve the decision making process of the physicians during the diagnosis of heart failure.

DEEPIKA et al. proposed predictive analytics to prevent and control all chronic diseases with the help of machine learning techniques such as naive Bayes, support vector machine, decision tree, and artificial neural network and they have used UCI machine learning repository datasets to calculate the accuracy. From the experiment, it is been found that SVM gives the highest accuracy rate of 95.556% in case of heart disease and the case of diabetes Naïve Bayes classifier gives the highest accuracy of 73.588%. Fitting predictive models like those utilized as a neighborhood of this study could be utilized to make individual/clinical decision support systems, to stay up wellness or enhance the administration of chronic diseases, for instance , heart condition and diabetes. Recognizable proof of real hazard factors and creating a choice network, and powerful control measures and wellness programs will decrease the chronic disease mortality

IV. Comparative Study of Literature Survey

No	Author	Techniques Used	Accuracy Obtained
1	LIAQAT ALI et al.	X^2 statistical model, deep neural network	93.33%
2	DR.KANAK SAXENA et al	Decision tree	86.3% (testing phase) 87.3% (training phase)
3	AWAIS NIMAT et al	Support vector machine, Hybrid grid Search algorithm (HGSA**)**	92.22% (L1 linear SVM+L2 linear & RBF SVM)
4	DEEPIKA et al	Naive Bayes, Decision tree, Support vector machine	SVM gives the best accuracy with 95.55%
5	ASHIR JAVEED et al	Random search algorithm (RSA), Random forest.	93.33% (RSA+RF)
6	D.M CHITRA et al	Decision tree, Support vector machine, Naïve Bayes.	Chronic disease diagnosis Between 82% and 92%

Table 4.1: Comparison of literature survey

V. Conclusion

Heart disease is a very critical issue in the present growing world. So, there is a need for an automated system to predict heart disease at earlier stages. So that it will be useful for the physician to diagnose the patients efficiently, and it will be useful to the people also because they can track their health issues by using this automated system. Some of the expert automated systems were summarized in this paper. Feature selection and prediction, two are essential for every automated system. By choosing features efficiently, we can achieve better results in predicting heart disease. We have summarized some algorithms which are useful while selecting the features, like hybrid grid search algorithm and random search algorithm, etc. So, in the future, it is better to use search algorithms for selecting the features, and then applying machine learning techniques for prediction will give us better results in the prediction of heart disease.

References

[1] L. Ali et al., “An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure,” IEEE Access, vol. 7, pp. 54007–54014, 2019, doi: 10.1109/ACCESS .2019.2909969.

[2] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, “An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection,” IEEE Access, vol. 7, pp. 180235– 180243, 2019, doi: 10.1109/ACCESS.2019.2952107.

[3] M. Gjoreski, A. Gradisek, B. Budna, M. Gams, and G. Poglajen, “Machine Learning and End-to-End Deep Learning for the Detection of Chronic Heart Failure from Heart Sounds,” IEEE Access, vol. 8, pp. 20313–20324, 2020, doi: 10.1109/ACCESS.2020.2968900.

[4] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, “An Automated Diagnostic System for Heart Disease Prediction Based on $\chi 2$ Statistical Model and Optimally Configured Deep Neural Network,” IEEE Access, vol. 7, pp. 34938–34945, 2019, doi: 10.1109/ACCESS.2019.2904800.

[5] M. R. Ahmed, S. M. Hasan Mahmud, M. A. Hossin, H. Jahan, and S. R. HaiderNoori, "A cloud-based four-tier architecture for early detection of heart disease with machine learning algorithms," 2018 IEEE 4th Int. Conf. Comput. Commun. ICCC 2018, pp. 1951–1955, 2018, doi: 10.1109 /CompComm . 2018. 8781022.

[6] “types of heart disease.” [Online]. Available: https://www.heartandstroke.ca/heart/what-is-heart-disease/typesof-heart-disease.

[7] J. Schmidhuber, “Deep Learning in neural networks: An overview,” Neural Networks, vol. 61, pp. 85–117, 2015, doi: 10.1016/j.neunet.2014.09.003.

[8] N. H. Farhat, “Photonic neural networks and learning machines the role of electron-trapping materials,” IEEE Expert. Syst. their Appl., vol. 7, no. 5, pp. 63–72, 1992, doi: 10.1109/64.163674.

[9] A. K. M SazzadurRahman, M. MehediHasan, S. Asaduzzaman, M. Asaduzzaman, and S. AkhterHossain, “An analysis of computational intelligence techniques for diabetes prediction Machine Learning View project An analysis of computational intelligence techniques for diabetes prediction,” Int. J. Eng. &Technology, vol. 7, no. 4, pp. 6229–6232, 2018, doi: 10.14419/ijet.v7i4.28245.

[10] G. H. Tang, A. B. M. Rabie, and U. Hägg, “Indian hedgehog: A mechanotransduction mediator in condylar cartilage,” J. Dent. Res., vol. 83, no. 5, pp. 434–438, 2004, doi: 10.1177/154405910408300516.

[11] Y. Karaca and C. Cattani, “7.Naive Bayesian classifier,” Comput. Methods Data Anal., pp. 229–250, 2018, doi: 10.1515/9783110496369-007.

[12] Purushottam, K. Saxena, and R. Sharma, “Efficient Heart Disease Prediction System,” ProcediaComput. Sci., vol. 85, pp. 962–969, 2016, doi: 10.1016/j.procs.2016.05.288.

[13] K. Deepika and S. Seema, “Predictive analytics to prevent and control chronic diseases,” Proc. 2016 2nd Int. Conf. Appl. Theor.Comput.Commun.Technol. iCATccT 2016, no. January 2016, pp. 381–386, 2017, doi: 10.1109/ICATCCT.2016.7912028.

[14] “Analysis and Prediction of Various Heart Diseases Using Dnfs Techniques,” vol. 2, no. 1, pp. 1–7, 2015.