Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020)
IEEE Conference Record #48766;IEEE Xplore ISBN: 978-1-7281-5371-1

Review on Heart Disease Classification

Harshali Dube[1], Shweta Madge[2], Prajakta Jagtap[3], Pooja Potdar[4], and Mr.Nilesh Bhandare[5]
[1,2,3,4] *Student, School of Computer Engineering and Technology, MIT Academy of Engineering, Alandi(D).*
[5]*Assistant Professor, School of Computer Engineering and Technology, MIT Academy of Engineering, Alandi(D).*
*Contact:[1] hrdube@mitaoe.ac.in, [2]ssmadge@mitaoe.ac.in,[3] prajaktajagtap@mitaoe.ac.in, [4] pdpotdar@mitaoe.ac.in,
[5] ndbhandare@mitaoe.ac.in

Abstract—**Heart disease is the foremost cause of death. According to WHO, deaths due to heart disease account for 30% of the global deaths. The main challenge that medical practitioners face is the disease not being identified at an early stage as the traditional approaches are quite time-consuming. Machine learning algorithms can help to deal with the emerging challenges in heart disease classification. The proposed work presents a short review about heart disease classification problem, many classification techniques along with future research directions.**

Keywords-Heart disease, classification techniques, support vector machine, genetic algorithm, GASVM.

INTRODUCTION

Heart disease is one of the leading reasons for loss of life. It has a harmful permanent disorder. Medical data has a vast amount of information but definite knowledge cannot be extracted out of it. Also, diagnosing patients accurately and on time becomes important. If patients are handled using previous data then definitely there is a possibility to extend the life of the patients. Blood vessel diseases, heart rhythm problems, and heart defects when an individual is born are a similar kind's disease. Nowadays, the global majority of death occurred due to heart disease as compared to other diseases. According to the report of the WHO, cardiac disorder is the main cause of loss of life in the world. Globally 30 % of death occurred due to cardiovascular disease. According to a survey in middle-level countries, 80% of deaths happened due to cardiovascular disease. The detection of heart problems is the more trivial task of medical researchers and accuracy is a major concern [7]. The University of California at Irvine (UCI) dataset attributes listed in Table I along with the description.

TABLE I
LIST OF ATTRIBUTES

Sr. No	Attribute Name	Description
1	**Age**	Ranges from 29 to 77
2	**Sex**	1 –Male, 0- Female
3	**Cp**	Chest Pain 1-Typical Angina, 2-Atypical Angina 3-Non- Anginal Pain 4-Asymptomatic
4	**trestbps**	Resting blood pressure Ranging from 94 – 200
5	**Chol**	Cholesterol, Ranging from 126 – 564
6	**Fbs> 120**	Fasting Sugar 1 = True; 0 = False
7	**restecg**	Resting ECG 0-Normal ECG 1-Having ST-T wave abnormality
8	**thalach**	Maximum Heart Rate Achieved Ranging from 71-202
9	**exang**	Exercise-Induced Angina 1-Yes0-No
10	**oldpeak**	ST depression induced by exercise relative to rest Ranging from 0 – 6.2
11	**slope**	The slope of the peak exercise ST segment, Ranging from 1-3
12	**ca**	major vessels Min 0- Max 3
13	**Target Variable**	0- No Disease 1- Disease Present

A. Classification

Various data mining techniques are used by the researcher. Classification is one technique that is mostly used to prepare various models. The basic objective of the classification is to divide the data into different classes and groups. In the process of classification, data will be scan according to class. In classification firstly find out the best classification model, which is compact and suitable to class.

B. Techniques and Methods

Naive Bayes :

Bayes' Theorem does the task of finding the probability of an event that is going to occur when the probability of another event that has previously happened is given. Mathematical model of the "Bayes' theorem" is described as below:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In the above theorem, Event A Probability will be calculated by Considering Event B has already happened. In this case, B event is the proof of event and B event considered as a hypothesis. Here some of the assumptions are involved i.e Predictor and features are not depends on each other. In this technique, no feature will be affected by other features so this process is called naive.

Authorized licensed use limited to: University of Exeter. Downloaded on July 15,2020 at 07:56:43 UTC from IEEE Xplore. Restrictions apply.

Generalized Linear Mode:

By the statistics, the GLM model is more generalization of Ordinary linear regression. The Ordinary linear regression allowed the response variable but those having error distribution models by holding normal distribution in the response variable. The mathematical model of GLM as below

$$E(Y)=\mu = g^{-1}(X\beta)$$

Where "E(**Y**) is nothing but outcome value for the Y, linear predictor-$\mathbf{X}\boldsymbol{\beta}$, unknown parameters linear combination $\boldsymbol{\beta}$; link function- g ".

Logistic Regression (LR):

LR is one of the classification algorithms exercised to assign observations to a fixed set of classes. LR changes its output by taking into account the logistic sigmoid function to return a probability value.

$$f(x) = \frac{1}{1+e^{-(x)}}$$

Deep Learning:

Deep learning inspired by artificial neural networks and the sub-algorithm of machine learning. Deep learning may be supervised, semi-supervised or unsupervised depending upon the condition

Decision Tree:

The decision tree looks like a flow chart diagram. It includes leaf nodes and each non-leaf node plays a test on each attribute and finally, each branch of the tree shows the result. As per other tree structures here also uppermost node will be considered as the root node of the decision tree. Although it handles multi-dimensional data, it faces repetition and replication issues. To improve its performance, attribute selection is utilized.

Random Forest and Gradient Boosted Trees:

Random forest is one more technique to solve classification and regression problems. Random forest work similarly like other classification techniques but the major difference is RF creates a decision tree for every attribute. Random forest deletes missing data values also delete extreme low and high values. The random forest creates dynamic models by using weak models. Example of random forest tree defined in figure 1.

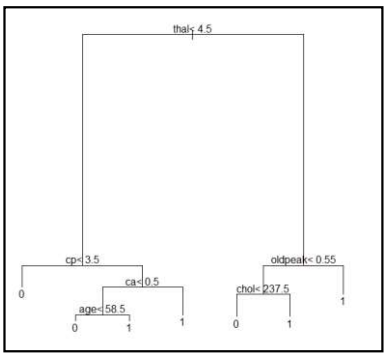

Figure 1: Random Forest Tree [19]

Neural network (NN)

NN is a chain of algorithms that endeavors to identify existing associations in a set of data via a procedure that is similar to how a human brain works. NN can adjust to changing input; so that the network generates the best result without redesigning the result specification. A "neuron" in a neural network can be defined as a mathematical function that collects as well as classifies information based on a specific structure. This network has a strong similarity to curve fitting and regression analysis.

SVM

A Support Vector Machine (SVM) is a classification technique and SVM classifier that is defined by a distinguishing hyperplane that discriminates two or more different classes. In simple words, SVM differentiates data in two classes in a very well manner. There is a large possibility to draw more than one hyperplane the data points. But most important is to find two hyperplanes with a maximum distance between them. SVM is more popular in its category because of the remarkable generalization ability of SVM. SLM belongs to the supervised learning category. Most of the researchers used SLM very effectively to build classification and regression models. Classification finds the classes and divides them into two or more than two classes. SVM create classifier which is non-probabilistic binary linear classifier [19]

K-nearest neighbors (KNN)

KNN is a simple machine learning algorithm that belongs to the supervised learning category and solves both classification and regression problems. KNN finds similarity between data points and assigns new data points depending upon similarity. According to that value of K will be calculated. So whenever a new data point will come it goes into a similar category.

LITERATURE REVIEW

In [1] the author proposed several classification techniques. Classification based techniques contribute high potency and obtain high correctness in contrast with the traditional method. The hybrid approach of support vector machine and ANN is the finest binary classification system for predicting heart disease [2]. Mostly, heart diseases are incurable by nature and these diseases make dangerous complexities. The SVM-ANN hybrid classifier approach gives an accuracy rate of about 88.54 % than the earlier proposed method [2]. SVM-ANN is

Authorized licensed use limited to: University of Exeter. Downloaded on July 15,2020 at 07:56:43 UTC from IEEE Xplore. Restrictions apply.

more precise than the single data mining algorithm. A hybrid random forest with linear model (HRFLM) is proposed in [3]. Before applying classification author used the selection model to select the attributes from the database. In [3] selection of attributes done by manually and achieved results compared all traditional machine learning algorithms like Naive, GLM, LR, DT, RF, and GBT. HRFLM achieved better results compared with simple data mining techniques by considering UC Dataset. SVM classification is compared with ANN classification [4] which is extremely empowering. The accuracy rate between SVM and ANN is noticeable.

The performance of the SVM classifier can be enhanced by ignoring repetitive search tasks for the specific end goal.

Connected to the errand of understanding the classification of heart disease data set based on account measurable values, here accuracy of SVM is higher. Possible to identify similar features by using correlation-based selection method and particle swarm optimization (PSO) / ant colony optimization (ACO) based optimization is utilized to optimize feature from the dataset to improve the result of heart disease classification [5]. Various machine learning algorithms are compared with the proposed hybrid approach. Merging of FCBF, Particle Swarm Optimization and ACO with KNN work very well and achieved 99.6 % results. It also obtained a 99.65 % result with RF. In [7], UCI Dataset used to create classifiers by using kernels. Here the author proposed two versions of reduction methods. These reduction methods called using names SVM-RFE, PCA-SVM. The author compared the obtained results with the help of these reductions techniques. SVM-RFE used to find the most important features from the dataset and it finds better performance in feature selections. Hybrid Differential Evolution based Fuzzy Neural Network (HDEFNN) is proposed to solve heart disease classification problems [7]. Also, DE (Differential Evolution) is fused with FNN (Fuzzy Neural Network). The proposed method obtained an accurate and reliable identification of heart disease with the help of NN-based learning. The system is said to achieve 69.1 % accuracy which was highest when compared with J48, Naïve Bayes, and Random Forest. In [8] two additional attributes are considered i.e. obesity and smoking. Simple data mining techniques are used to analyze heart disease classification. NN predict better result than Naive Bayes and Decision Tree. The collective approach of CANFIS (Coactive Neuro-Fuzzy Inference System) and Genetic algorithm is used to predict heart disease classification [10]. The CANFIS model integrates fuzzy inputs with a modular neural network to \ approximate complex functions accurately and fast. A genetic algorithm is applied to identify the most suitable features from the dataset which produces a smaller and less complicated network and removes redundant variables. Identification of the best features set and CANFIS parameter's auto tunning is done through optimizing GA in very effectively. Multi-layered feedforward networks are used to solve the cardiovascular disease prediction problem. Multi-layered feed-forward networks applied to the UCI dataset. In that 303 records are present which exist in 14 class attributes. Genetic algorithms find good feature selection to problems quickly and neural networks are adaptive models for data analysis. The accuracy achieved is 94.17 % [11]. To identify suitably and appreciate features for the classification and prediction purpose author take the help of an evolutionary genetic algorithm. [12][16]. The accuracy of the different traditional algorithm is described in figure-2.

Support Vector Machine used along with the Genetic Algorithm to predict Blood-Brain Barrier Penetration [13]. SVM model prepared by using the chromosomes which contain different parameters. SVMs have supervised learners that build a model from training data with known classification [14] SVM parameters are required to set using the requirement of the model to find accurate results from the models because SVM gives different parameters. The exact combination of this parameter required to prepared a good model.

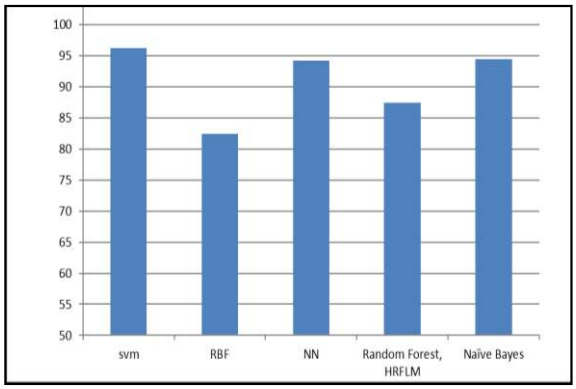

Figure 2: Accuracy Comparison

Model selection for SVM classifier done through genetic algorithm [14]. There however exist some issues with such an approach like reliability. The issues are further handled in the next research. SVM result also depends on parameters setting. The parameter setting is nothing but a model configuration. In [15] popular genetic algorithm used very effectively to identify the best configuration of required parameters for the prediction and classification. SVM is a widely used classification and prediction technique. However, there are some limitations with using SVM so the author introduced a solution to identify error-prone components/parts, using the genetic algorithm (GA) and SVMs. GA is used to search for a suitable SVM parameter setting that exploits fitness function. Cancer cell classification problem deal with GA and SVM Firstly genetic algorithm applied to search suitable feature set and then cell classified into the healthy cell and cancel cell by using classification techniques [16]. It is possible to deal with linear and non-linear problems classification with the help of supervised SVM.

In [17] the author used genetic heuristics to optimize the results with the help of SVM. Genetic used to find optimize features from the dataset. The accuracy obtained with the combination of genetics and SVM is 79.74%. In [18], the author compares the classification accuracy of the two approaches. These include the optimized SVM which is based on quantum genetic algorithm and the conventional SVM with grid search. The obtained accuracy is 93.85%. The proposed approach shows better performance with an accuracy of 96.15%.

Authorized licensed use limited to: University of Exeter. Downloaded on July 15,2020 at 07:56:43 UTC from IEEE Xplore. Restrictions apply.

TABLE III
FEATURE SELECTION USING GA FOR VARIOUS PROBLEM

Sr.	Problem	Approach	Remark
1	Heart Prediction System.[10]	CANFIS& GA	CANFIS integrates adaptable fuzzy logic and GA helps in feature selection
2	Prediction System[11]	GA &NN	Multi-layered feed-forward networks for optimal results.
3	Enhanced classification[12]	GA & Naïve Bayes, DT	It helps to eliminate irrelevant attributes and giving the best possible combination of features.
4	Blood-Brain Barrier PenetrationPrediction[13]	GA &SVM	GA used to search for suitable configurations of SVM parameters.
5	Predicting Fault-Prone components[15]	GA	Identification of error-prone components using a proper fitness function.
6	Time Series Classification[17]	GA& SVM	Controls overfitting by refining the features
7	Cancer Detection[16]	GA& SVM	GA used for feature selection and SVM to classify whether healthy and cancerous cells
8	Human Action Recognition[18]	Conventional SVM + Quantum GA	Optimized results using hybridization of SVM and Quantum GA.

PROPOSED WORK

A. Data Pre-processing

In data pre-processing step raw data converted into readable data by using different data pre-processing steps. The processes that fall under data pre-processing are cleansing, editing, reduction, and wrangling of the data in various ways. The pre-processing that carried out includes removing the duplicate values & removing NA (null) values.

B. Feature Selection

Feature selection is nothing but the process of finding and ignoring irrelevant, repeat attribute, or less required attributes or some time dimensions from the dataset.
UCI dataset contains 13 attributes or features. The evolutionary genetic algorithm applied to identify suitably and appreciate features for classification and prediction purposes.

C. Genetic Algorithm

A genetic algorithm is belonging to the evolutionary algorithm category, identified by "Charles Darwin's theory of natural evolution". In GA firstly random population will be created and after applying fitness function individuals send to the reproduction step. It is possible to extract the best features with the help of a genetic algorithm. Here firstly need to create an initial population that is nothing but a different combination of features of datasets. After that, the fitness value of each chromosome needs to calculate using fitness functions. Finally need to apply Genetic algorithm iteratively till stopping criteria. The flow chart (figure 3) shows the working of a genetic algorithm.

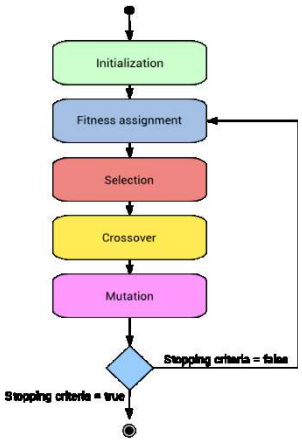

Figure 3: GA's Flowchart

D. SVM

The SVM is a major technique for the classification of jointly linear and non-linear data. SVM gives better accuracy compared to other traditional algorithms. In SVM classification, the classifier identifies hyperplane which finds the maximum distance between two classes while minimizing the classification errors.

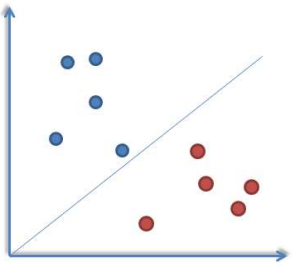

Figure 4: SVM

E. Proposed *GASVM* to predict Heart Disease

The proposed GA+SVM approach works as follows. The Optimized Genetic Algorithm is applied to identify the suitable combination of the features of attributes from the dataset after the data pre-processing step. The genetic algorithm will remove the redundant features from the dataset and gives the best feature combination to making the prediction accuracy better. Further support vector machine is used for the classification and prediction of heart diseases. Proposed GASVM architecture shown in figure 5.

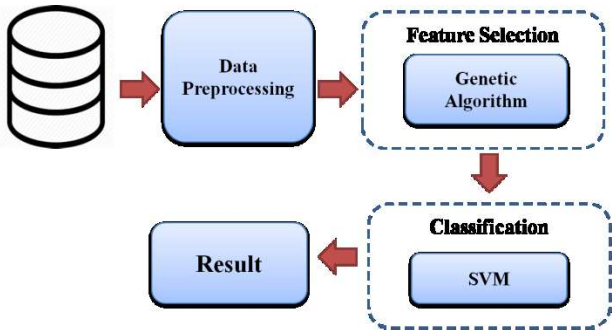

Figure5: Proposed Architecture GASVM

Authorized licensed use limited to: University of Exeter. Downloaded on July 15,2020 at 07:56:43 UTC from IEEE Xplore. Restrictions apply.

CONCLUSION

Heart disease is the major cause of death all around the globe. Due to increasing complexities, it is the need of the time that the diagnosis takes place faster so that patients can be treated well. Traditional approaches are time-consuming and accuracy is a major concern. The primary challenge is to predict accurately and faster. Machine learning techniques can help with this concern. Some of the attributes are not directly required to predict heart diseases. The genetic algorithm will remove the redundant and unnecessary features from the dataset and give the best feature combination to making better prediction accuracy. SVM gives better accuracy compared to other traditional algorithms with GA. This paper illustrates that it may be achievable to upgrade the accuracy of cardiovascular disease by using GA for feature selection and SVM for Classification.

REFERENCES

[1]. C. Sowmiya and P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques," *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Srivilliputhur, 2017, pp. 1-5.*

[2]. C. Yang, B. An and S. Yin, "Heart-Disease Diagnosis via Support Vector Machine-Based Approaches," *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Miyazaki, Japan, 2018, pp. 3153-3158.

[3] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in *IEEE Access*, vol. 7, pp. 81542-81554, 2019.

[4] S. Radhimeenakshi, "Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network," *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2016, pp. 3107-3111

[5] Khourdifi, Youness and Mohamed Bahaj. "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization." (2019).

[6] Yang, Chengming& An, Baoran& Yin, Shen. (2018). Heart-Disease Diagnosis via Support Vector Machine-Based Approaches. 3153-3158. 10.1109/SMC.2018.00534.

[7] Mohan, Senthilkumar&Thirumalai, Chandra Segar & Srivastava, Gautam. (2019). Effective Heart Disease Prediction using Hybrid Machine Learning Techniques. *IEEE Access*. PP. 1-1. 10.1109/ACCESS.2019.2923707.

[8] Dangare, Chaitrali S. and S. S. Apte. "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques." (2012).

[9] Ramalingam, V V&Dandapath, Ayantan& Raja, M. (2018). Heart disease prediction using machine learning techniques: A survey. International Journal of Engineering & Technology. 7. 684. 10.14419/ijet.v7i2.8.10557.

[10] Parthiban, Latha& Subramanian, R.. (2007). Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm. *Intl J Biol Life Sci.* 3.

[11] N. G. B. Amma, "Cardiovascular disease prediction system using genetic algorithm and neural network," 2012 *International Conference on Computing, Communication and Applications*, Dindigul, Tamilnadu, 2012, pp. 1-5.

[12] Masilamani, Anbarasi& ANUPRIYA, & Iyenger, N ChSrimanNarayana. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. *International Journal of Engineering Science and Technology*. 2.

[13] Zhang, Daqing& Xiao, Jianfeng& Zhou, Nannan&Zheng, Mingyue&Luo, Xiaomin& Jiang, H. & Chen, Kaixian. (2015). A Genetic Algorithm Based Support Vector Machine Model for Blood-Brain Barrier Penetration Prediction. BioMed Research International. 2015. 10.1155/2015/292683.

[14] S. Lessmann, R. Stahlbock and S. F. Crone, "Genetic Algorithms for Support Vector Machine Model Selection," *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, Vancouver, BC, 2006, pp. 3063-3069.

[15] Di Martino S., Ferrucci F., Gravino C., Sarro F. (2011) A Genetic Algorithm to Configure Support Vector Machines for Predicting Fault-Prone Components. In: Caivano D., Oivo M., Baldassarre M.T., Visaggio G. (eds) *Product-Focused Software Process Improvement. PROFES 2011. Lecture Notes in Computer Science,* vol 6759. Springer, Berlin, Heidelberg

[16]Mansoori, Khan & Suman, Amrit& Mishra, Sadhna. (2014). Application of Genetic Algorithm for Cancer Diagnosis by Feature Selection. *International Journal of Engineering Research & Technology*. Vol. 3 Issue 8. 1295-1301.

[17] W. K. Resende, R. A. Nascimento, C. R. Xavier, I. F. Lopes and C. N. Nobre, "The use of support vector machine and genetic algorithms to predict protein function," 2012 *IEEE International Conference on Systems, Man, and Cybernetics (SMC), Seoul*, 2012, pp. 1773-1778.

[18] Liu, Yafeng et al. "Highly Efficient Human Action Recognition with Quantum Genetic Algorithm Optimized Support Vector Machine." ArXiv abs/1711.09511 (2017)

[19] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, 2018, pp. 1-7.

Authorized licensed use limited to: University of Exeter. Downloaded on July 15,2020 at 07:56:43 UTC from IEEE Xplore. Restrictions apply.