Research Article

Predicting Heart disease using Machine Learning

Nitant[a], Dr. Rashmi Priya[b]

[a]Phd scholar, G D Goenka University
[b]Assistant Professor G D Goenka University

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract: in recent years a spike is seen in mortality linked to heart disease. In developed countries mortality linked to heart diseases have seen a gradual decrease due to the availability of better medical facilities and public awareness but on the other hand in developing countries like India and other south Asian countries alarming increment of mortality linked to heart disease has been seen in recent years. According to a research published in Indian heart journal India has seen a 4 fold increment in mortality linked to heart disease in past 40 years [1]. Abundant amount of data is collected by various medical institutions in past a decade. The data is available on various open source platforms like UCI, KAGGLE, and other Government websites.in this paper we will review all the techniques of data mining, machine learning and deep learning used by the researches previously on these open source datasets. So that we can propose a better technique for early prediction of heart disease using machine learning and deep learning models.

Keywords: heart disease data set, decision tree model, artificial neural network model.

1. Introduction

The Knowledge is the key component of intelligence. Knowledge is built by learning. While developing a better Artificial The intelligence (AI) model we require knowledge which is obtained by the various algorithms of machine learning and deep learning, hence machine learning is often termed as sub field of AI. In today’s digital technology dependent era humans have more dependent on the machines for the basic works and thus resulting in less physical work, Due to less physical work there is a never seen increase in health issues like obesity, fatigue which are the main driving force behind Coronary heart disease. Technology is used for building a better world by introducing wearable gadgets through which the health related data is collected and used for research purpose to identify the health issues at an early stage. The domain of Machine learning has gained much more importance in the field of the medical sector for solving the clinical problems.

Majorly found heart related disease is the Coronary heart disease (Coronary Artery Disease). The blood flows through coronary arteries in the heart it gets narrow due to the formation of the atherosclerotic plaques which leads to heart attack clinically known as Myocardial Infarction. The best part of Coronary heart disease is that it can be revised with proper medication if detected at an early stage, for detection already we have many pathology based test. Coronary angiography is most followed standard method for detection of coronary heart disease, but due to its invasive nature and high cost it cannot be done frequently .other available pathology based methods are exercise electrocardiogram which is now available on the wearable gadgets, stress echocardiography, single photon emission computed tomography , electron-beam computerized tomography. By using the data from the above techniques it is easier to detected coronary heat disease. In ECG the coronary heart disease is detected by observing the changes in the ST segments or twaves. The coronary heart disease can be related to any of the four values of the heart for detecting them the MRI scan is available which develops a 3d model rendering of the heart. The various factors for development of Coronary heart disease are less physical activity (due to immense use of technology all the services are available to the door steps and thus making us more lazy), high blood pressure which result in excessive blood flow to the heart and making the walls of the vessels to overstretch and causes injuries, family history of Coronary heart decease it is possible that it gets inherited from family, smoking a smoker is 40% more likely to develop coronary heart disease as compared to non-smokers, cholesterol is a lipid fat which is waxy in nature which flows in the blood ,increase in cholesterol is also a major reason for development of coronary heart disease, obesity is a condition where the patient has more weight above the ideal body weight is also lined to Coronary heart disease.

Classification of Coronary heart disease has been one of the major topic for researchers. Researches have already used various data mining, machine learning based techniques for the classification of Coronary heart disease. Majorly the neural network of machine learning field has been used for the development of models for the classification, apart from neural networks many supervised and unsupervised machine learning techniques are also used.

2. Literature review

Many research have been done for the classification of Coronary heart disease by using the machine learning techniques. For training the machine learning model the data set is available on the uci machine learning repository. Mostly used data set is the Cleveland dataset. The Cleveland data set contains 303 records with 14 attributes namely age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina, old peak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) coloured by fluoroscopy , thal: 3 = normal; 6 = fixed defect; 7 = reversible defect, num (the predicted attribute). Originally the data set contains 72 attributes out of which these 14 attributes are commonly used by the researchers.

Sharma Purushottam et al

Sharma Purushottam et al, [2] proposed a c45 rule (decision tree) and partial tree methodology for heart disease prediction .the dataset used for this paper is the Cleveland Clinic dataset which is available on the UCI .many rules were discussed in this paper for building a decision tree with the help of a hill climbing algorithm and exhaustive algorithm. The parameters used for hill climbing algorithm were confidence (minimal confidence by a node to be considered as a node in the tree, value used was 0.25), minItemsets (minimum number of leaf under a node, value used was 2) and threshold (it was used for finding best subset under a node if the value is less than threshold then an exhaustive algorithm is used or else hill climb algorithm is used, value is 10). The paper was implemented using KEEL and WEKA tools. The result for the proposed methodology showed accuracy of 86.7% compared to the accuracy of decision tree which was 73.5.limitation of the research is that the data used for training the model was not treated for skewness and missing values. The data with class imbalance (skewness) would result in over fitted or under fitted model. For result and analysis K-Fold cross validation was not used. The missing values were replaced by using the mean of the values.

YounessKhourdifi et al

YounessKhourdifi et al, [3] proposed a Particle Swarm Optimization (PSO) combined with Ant Colony Optimization (ACO) and Fast Correlation-Based Feature Selection (FCBF) based optimized machine learning classifier. The data set used was the UCI heart disease data. WEKA tool was used for the implementation. The classification models used were K-Nearest Neighbors, Support Vector Machine, Naïve Bayes, Random Forest and a Multilayer Perception | Artificial Neural Network optimized by PCO and ACO. The best model were K-Nearest Neighbors and Random Forest after the optimization. Fast Correlation-Based Feature Selection was used for preprocessing of the dataset which included dimension reduction and removing the redundant data. the accuracy achieved by each classifier after the optimization were K-Nearest Neighbors of 99.6%, SVM 83%, Random forest 99.7%,NB 85%,MLP 90%. Limitation of the research is that it used uci Cleveland dataset which is skewed and contains the missing values. The skewed dataset is used for the feature selection process using ant colony optimization, no such work is done on the part of the machine learning algorithm. For selecting the features the ant colony optimization divides the data in to multiple smaller units and then using the correlation groups are formed. Due to class imbalance there is a greater chance of overfitting of machine learning models that were used including SVM, KNN, Logistic regression and mlp.

Costa W.L et al

Costa W.L et al, [4] proposed a Decision support system using artificial neural network (ANN).the ANN used consist of a single hidden layer. Hidden layer consisted of 6 neurons and sigmoid as the activation function with a learning rate of 0.28, the optimization was done by using Nesterov’s momentum optimizer to find the global minimum. The dataset used for this paper is the Cleveland Clinic dataset which is available on the UCI Machine Learning repository. The model has an accuracy of 90.76% and precision, recall and f1score to be 0.91. The Amazon EC2 cloud server was used for the hyper parameter tuning of the model. Limitation of the research is that the skewness of the data was not removed and the missing values were replaced by using the median of the values. The model was built by hyper tuning the parameters. The sigmoid function was used with a learning rate of 0.28. The sigmoid function would get struck in a local minimum and hence cannot reach an optimum value. Further due to hyper parameter tuning the model could not work better for new data as it is trained for the specific pattern.

Amarbayasgalan T et al

Amarbayasgalan T et al, [5] proposed a Deep Auto encoder based neural network for predicting heart disease. KNHANES datasets was used for the research. KNHANES datasets is available on Korea National Health & Nutrition Examination Survey website. Auto encoder is a type of artificial neural network (ANN) which is categorized in to unsupervised learning used for the dimension reduction. The auto encoder used consisted of three hidden layers where 4, 1, 4 neurons were used respectively. The neural network with 5 hidden layer is used with 17, 9, 5, 3, and 2 nodes, respectively. The ReLu activation function is applied to each hidden layers and the sigmoid activation function is applied to the output layer. Also, Adam optimizer which is a stochastic gradient based optimizer is used for weight optimization. The accuracy of model is 83.53%, precision 89.56%, F-measure 84.36% and AUC score 84.02%. The results were generated using sci-kit learn open source library of python. Limitation of the research is that it uses the KNHANES data set which contains missing values and the data is skewed. Deep auto encoders will partition the data set into two category with high variance and normal data set. But the skewness (data imbalance) was not addressed and further a neural network with relu action function was used with improper configurations. The last layer was of sigmoid activation is used which degrades the performance by considering all the input features. By considering all the input features the model will be under fitted.

K. G. Dinesh et al

K. G. Dinesh et al, [6] used the machine learning based algorithms for the prediction of heart disease. The dataset used for this paper is the Cleveland Clinic dataset which is available on the UCI Machine Learning repository. The models used were Logistic regression, naïve Bayes, random forest, gradient boost, svm. The data preprocessing included handling missing values, scaling data using standard scalars. For the implementation R language is used. The Logistic regression model has the highest accuracy among other models. Resultant accuracy of logistic regression is 86.51, random forest is 80.89, Naïve Bayes is 84.26, gradient boost is 84.26, and svm is 79.77. Limitation of the research is that it did not consider treating the class imbalance in the data. After processing the data the machine learning models like svm, logistic regression, were applied but due to the skewness the model were under fitted degrading the prediction accuracy. Further no such work was done on hyper tuning the parameters and no methods were employed for the feature selection.

Senthilkumar Mohan et al

Senthilkumar Mohan et al, [7] proposed a hybrid machine learning technique for the prediction of the Coronary heart disease. The model used decision tree for the process of the feature selection, for classification of the Coronary heart disease a hybrid model of random forest and linear model was built the linear model used was linear SVM. The data was spitted into multiple portion by using the decision tree then the prediction model was used and the error rate was calculated further the portion with the minimum error is selected.it uses the uci ml heart disease data set. The implementation was done by using the R language. The model resulted an accuracy of 88.4%. Limitation of the research was it did not consider the treatment of the skewness and missing values of the data. The attributes like age and sex were not used during the feature selection process. The entropy was used for the information gain in the decision tree model which is logarithmic in nature and takes more time as compared to the GINNI index method.

Norma LatifFitriyani et al

Norma LatifFitriyani et al, [8] proposed an effective heart disease prediction system using the xgboost classifier. The uci dataset is used for building the model. The dataset pre-processing includes removal of the outliers by using the DBSCAN clustering. After finding the optimal eps for the DBSCAN the clustered data is used and the outliers are left behind. As we know that the uci data is skew ed so for removing the data skewness synthetic over sampler (SMOTEENN) is used. The model resulted an accuracy of 95.9%. The python language along with Xgboost open source library and scikit learn open source library is used for the implementation. The limitations of the research are that it does not have any process for the feature selection and for the result analysis the K-fold cross validation is not used. The accuracy presented is the highest achieved accuracy of the model not the average of many trails.

Jian Ping Li et al

Jian Ping Li et al, [9] proposed a novel KNN based feature selection process. The uci data is scaled by using min max scaler and standard scaler then the proposed ‘FCMIM’ (fast conditional mutual information) algorithm is used for the feature selection .FCMIM algorithm assigns a local weight to each attribute and uses the KNN based algorithm to find the best input features. The selected features were used to train and test various machine learning classifiers like Logistic regression, SVM, KNN, ANN and Decision tree. For better result analysis and building model LASSO cross validation is used. A highest accuracy of 85% was achieved by linear SVM classifier. Limitation of the work is the Low prediction accuracy which could be due to not processing the skewness of the data.

ApurbRajdhan et al

ApurbRajdhan et al, [10] proposed heart disease classification model by using the uci heart disease dataset.

The data was not processed for skewness and missing values. The approach used machine learning classifiers like Logistic regression, random forest, Naïve Bayes and decision tree which resulted accuracy of 82, 90, 86 ,85 respectively. The main limitations are no proper pre-processing of data and low prediction accuracy.

Devansh Shah et al

Devansh Shah et al, [11] compared various machine learning classifiers for prediction of heart disease .KNN algorithm achieved a highest accuracy of 90% and rest algorithms like Naïve Bayes, Decision tree, random forest resulted accuracy of 80%, 81%, 82% respectively. The uci dataset is used for training the various models.The main limitations are no proper pre-processing of data and low prediction accuracy.

Rajesh N et al

Rajesh N et al, [12] compared Naïve Bayes and Decision tree machine learning classifiers to build a prediction model for the classification of heart disease. The Naïve Bayes algorithm worked better on the small portion of dataset and Decision tree for large dataset. The uci data set was used for building the model but the data was not processed for skewness and missing values. The work was implemented by using the R language platform. The prediction accuracy was low.

3. Summary of the literature review

reference	technique	limitation	Advantage	accuracy
Sharma Purushottam et al	Decision tree with the help of a hill climbing algorithm and exhaustive algorithm.	Low prediction accuracy	Better information gain by using the c 4.5 rules and hill climb algorithm	86.7%
YounessKhourdifi et al	PSO + ACO + FCBF	Over fitted model due to Skewness of the dataset	Better and faster feature selection by using the Fast Correlation-Based Feature Selection (FCBF) High accuracy	99.7%
Costa W.L et al	Artificialneuralnetwork (ANN).	Computationally complex	High accuracy	90.7%
Amarbayasgalan T et al	Deep Auto encoder	Low accuracy	Low computation time	83.53%
K. G. Dinesh et al	Logistic regression + SVM	Computation time is high	Logistic regression resulted a high accuracy	86.51
Senthilkumar Mohan et al	Linear SVM +Random forest	Computationally complex and needs more time for execution	Minimizing the error to achieve high accuracy	88.4%
Norma LatifFitriyani et al	SMOTEENN + DBSCAN + XGboost	More execution time required to build model	High accuracy	95.9%
Jian Ping Li et al	FCMIM+LASSO CV + SVM	Computationally complex and low accuracy	A better feature selection process to improve the prediction accuracy	85%
ApurbRajdhan et al	Logistic regression + SVM + Random forest +Naïve Bayes	The accuracy of LR and SVM is low.	Random forest achieved a high accuracy	90%
Devansh Shah et al	KNN + SVM + Random forest +Naïve Bayes	Unhandled missing values and skewness of the data set.	KNN achieved a high prediction accuracy	90%
Rajesh N et al	Naïve Bayes + Decision	Unhandled missing values and skewness	NB worked better on small portion of data and	NA

	tree	of the data set. High computation time	decision tree worked better on larger dataset	

Table 1: Summary of the literature review

4. Future research

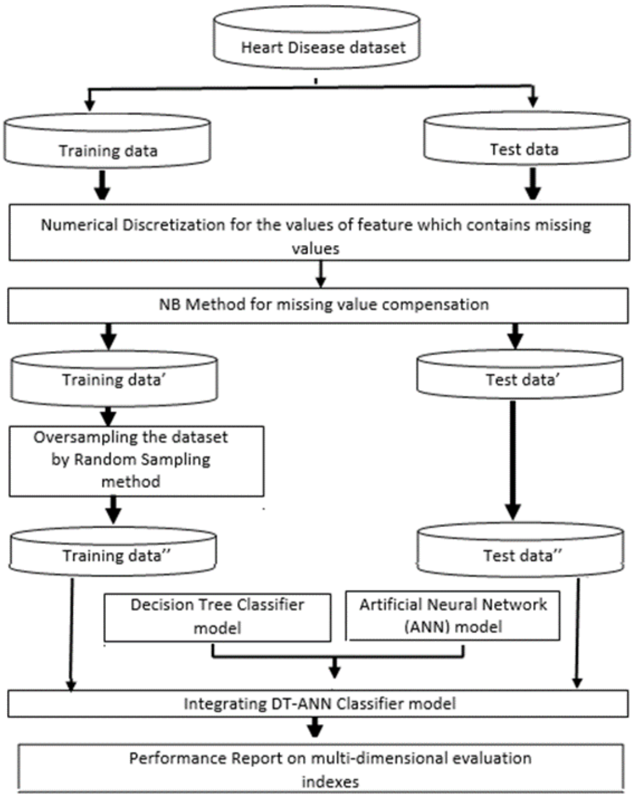
Heart Disease dataset
Training data
Test data
Numerical Discretization for the values of feature which contains missing values
NB Method for missing value compensation
Training data'
Test data'
Oversampling the dataset by Random Sampling method
Training data''
Test data''
Decision Tree Classifier model
Artificial Neural Network (ANN) model
Integrating DT-ANN Classifier model
Performance Report on multi-dimensional evaluation indexes

Fig 1. Proposed Model for Predicting Heart disease using Machine Learning

Heart disease data set is available on the uci machine learning repository and the kaggle.com website also. Firstly we will divide the data in to training and testing samples by shuffling the data set. The training part of the data is used for model building aka training the model. The data may have missing values so we treat the missing values. For treating the missing values we will use Naïve Bayes model. Naïve Bayes is a probabilistic model based on the Bayes' theorem [13].

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

Where A and B are events and B!=0. After treating the missing values we use the Pearson correlation coefficient to determine the correlation between the output and the input columns so that we could wisely choose the important columns in the feature set. The correlation coefficient is visualized by using a heat map. The training data is skewed that is imbalanced .hence we use the Random Sampler for resampling of the data. In oversampling we try to add more data in the training set for the minority class and in the under-sampling majority class data is removed to balance the data set. Random oversampling involves the process of replicating the row of the minority class dataset so that a balance can be brought. The oversampling will provide a better understand ability of the model regarding the entire population of the data set .rather than taking a portion of data from the population for building the model. The resampled training data is used for building the model. Artificial neural networks consist of multiple hidden layers each layer is used to implement some of the mathematical/statistical manipulations on the data the result of the layers is served as input to the next layers. Artificial neural networks are used to build a complex model with great accuracy, but on the other hand, it is difficult to understand or interpret. Due to the complexity, the artificial neural networks are not widely used in many fields. The training data set is used to build an artificial neural-network decision tree algorithm, which extracts binary decision trees from a trained neural network. The artificial neural networks with the Decision Tree algorithm use the neural network to generate outputs for samples interpolated from the training data set. In contrast to existing techniques, artificial neural networks with Decision Tree can extract rules from feedforward neural networks with continuous outputs. These rules are extracted from the neural network without making assumptions about the internal structure of the neural network or the features of the data. Schmitz [14] proposed an ANN-DT with a novel method for the attribute selection which can be used for the attribute selection. The attribute selection method is called as Weighted Variance Minimization. If a set consists of exemplars with corresponding output values, an input value and a threshold are required in order to split the data set into two sets and . For a least squares error measure the attribute and threshold are selected which because the maximum decrease in the weighted variance over the two branches, where the outputs of the data set are. In this way the sum of the squared errors of the newly formed branches is reduced with each split .This is the same procedure as is used in the CART algorithm when forming a regression tree. Attribute selection criterion based on a significance analysis of the variables on the neural-network output is examined. The test data set is used for the model evaluation. The result is analysed by using the confusion matrix, precision, recall, accuracy.

5. Conclusion

In our work we have proposed an ANN-DT (Artificial neural network decision tree) based model for predicting the heart disease. Previously various Machine learning, Datamining and Neural Network models were employed for the same. Machine learning and the Datamining did not have a good method for the selection of the important features. We will use the Decision tree as the activation function in the layers of the ANN so that we can have the important features out of the input in each layer and served as the input to the next layer of the ANN model. The ANN-DT approach can outperform the existing techniques for the heart disease prediction.

References

1. M.N. Krishnan. (2012). Coronary heart disease and risk factors in India – On the brink of an epidemic?. Indian Heart Journal. 64(4), 364-367. https://doi.org/10.1016/j.ihj.2012.07.001.
2. prushottama, Prof. (Dr.) KanakSaxena, Richa Sharma. (2016). Heart Disease Prediction System. Procedia Computer Science. 85, 962-969. https://doi.org/10.1016/j.procs.2016.05.288.
3. YounessKhourdifi, Mohamed Bahaj. (2019). Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. International Journal of Intelligent Engineering and Systems, 12, 242-252. 10.22266/ijies2019.0228.24.
4. W. L. Costa, L. S. Figueiredo, and E. T. A. Alves. (2019). Application of an Artificial Neural Network for Heart Disease Diagnosis. IFMBE Proceedings. 70(2), 753-758. https://doi.org/10.1007/978-981-13-2517-5_115
5. Amarbayasgalan T., Lee J.Y., Kim K.R., Ryu K.H. (2019). Deep Autoencoder Based Neural Networks for Coronary Heart Disease Risk Prediction". In: Gadepally V. et al. (eds) Heterogeneous Data Management, Polystores, and Analytics for Healthcare. DMAH 2019, Poly 2019. Lecture Notes in Computer Science, vol 11721. Springer, Cham[online], https://doi.org/10.1007/978-3-030-33752-0_17
6. Dinesh Kumar G , Santhosh Kumar D , Arumugaraj K, K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari. (2018). Prediction of Cardiovascular Disease Using Machine Learning Algorithms. International Conference on Current Trends towards Converging Technologies (ICCTCT). 1-7. doi: 10.1109/ICCTCT.2018.8550857.
7. Senthilkumar Mohan, handrasegar Thirumalai1, And Gautam Srivastava. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. 7, 81542-81554. doi: 10.1109/ACCESS.2019.2923707.
8. Norma LatifFitriyani, Muhammad Syafrudin, GanjarAlfian. (2020). HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System. IEEE Access. 8, 133034-133050. doi: 10.1109/ACCESS.2020.3010511.
9. Jian Ping Li, AminUlHaq, Salah Ud Din,Jalaluddin Khan, Asif Khan, and AbdusSaboor. (2020). Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. IEEE Access. 8, 107562-107582. doi: 10.1109/ACCESS.2020.3001149.
10. ApurbRajdhan, Avi Agarwal, Milan Sai, Dundigalla Ravi, Dr. PoonamGhuli. (2020). Heart Disease Prediction using Machine Learning. International Journal of Engineering Research & Technology (IJERT). 09(04), 659-662. http://dx.doi.org/10.17577/IJERTV9IS040614.
11. Devansh Shah, Samir Patel, Santosh Kumar Bharti. (2020). Heart Disease Prediction using Machine Learning Techniques. SN Computer Science. 1. 10.1007/s42979-020-00365-y.
12. Rajesh N, T Maneesha, ShaikHafeez, Hari Krishna. (2018). Prediction of Heart Disease Using Machine Learning Algorithms. International Journal of Engineering & Technology, 7(2.32), 363-366, 10.14419/ijet.v7i2.32.15714.

13. Wikipedia, Bayes' theorem [online]. Available: https://en.wikipedia.org/wiki/Bayes%27_theorem
14. Young Joong Kim, Muhammad Saqlian, Jong Yun Lee. (2019). Deep learning–based prediction model of occurrences of major adverse cardiac events during 1-year follow-up after hospital discharge in patients with AMI using knowledge mining. PersUbiquit Computing, https://doi.org/10.1007/s00779-019-0124 8-7.