BioGecko Vol 12 Spl Iss 01 2023

ISSN NO: 2230-5807

A COMPARISON AND ANALYSIS OF SUPERVISED MACHINE LEARNING ALGORITHMS TOWARDS ACCURATE PREDICTING OF HEART DISEASES

Dr. Prashant Sharma [1],Avantika Mahadik [2], Dr. Vaibhav Narawade [3]

[1]Associate Professor, Department of Computer Science and Engineering,
Pacific (PAHER) University Udaipur, Rajasthan, India
[1] prashant.sharma@pacific-it.ac.in

[2]Research Scholar,Pacific (PAHER) University Udaipur, Rajasthan, India
[2]avantika_mahadik@rediffmail.com

[3]Professor, Department of Computer Engineering,Ramrao Adik Institute of Technology
[3]vaibhav.narawade@rait.ac.in

Abstract: In machine learning, forecasting is one of the most significant applications. Machine learning uses various techniques for prediction. Taking into consideration the recent work, we are focusing on machine learning algorithms and analysing how these algorithms are used in the healthcare industry to predict heart diseases. In supervised machine learning, the machine first states the patterns from labelled dataset (training dataset) and secondly it applies that on the unlabelled dataset (testing dataset) to predict the result. The training dataset includes input and correct output. Classification and Regression are the two techniques used in supervised machine learning. Classification technique is very commonly used to predict diseases in healthcare. Classification is a learning technique in which, deciding of class label to a given data done through machine learning algorithms. Regression technique shows the relationship between two or more variables. We discover links between dependent and independent variables through regression techniques.
The core focus of this exploration is to conduct a systematic proportional study and examine of four machine learning algorithms, specifically random forest, support vector machine, KNN and decision tree in heart disease prediction. We found that support vector machine and random forest provides the highest accuracy in the prediction of heart disease among all. Random forest can be integrated with another classifier to achieve more efficiency.

Keywords: Machine learning, random forest, decision tree, Support vector machine, KNN and Heart Disease,

Introduction:
Nowadays, various industries in public and private sectors generatevast quantity of data. Data is a key aspect and plays a vital role in new development. We need reliable information to produce optimized and best outcomes in any area. Data in an appropriate structure, organized manner and in a suitable predefined model is called structured data. Unstructured or raw data tend to give rise to several problems while working with them, especially if that data is used in analysis. Today healthcare industry is one of the biggest industries which has a huge amount of medical data. This industry collects the data from various sources like hospitals, insurance companies, pharmaceutical industries,Epidemiological Surveillance, census, other health records, sample registration system, patient's disease registries, electronic health record and clinical surveys. The collected data is stored and used for analysis for better improvement in the medical field, research in drugs and predication or diagnosis of disease. Medical data comes in many forms. Text, images, sound,various readings from wearable medical IoT devices, biosensors data, data collected from clinical instruments and devices

are forms of medical data. Still, we are challenged with managing, storing and analysing large amount of medical data. The occurrence of missing values is another issue that needs to be tackled. To learns, predicts and improves automatically based on previous experiences is a main task of machine learning. A machine learning algorithm uses classification and regression techniques. It gives thousands of models and predicts the outcomes. By providing with enormous patterns, it becomes easy to predict the diseasesand diagnose them atapremature stage. Machine learning algorithms are commonly used to predict diseases like thyroid, diabetes, heart diseases and some brain related diseases in the area of medicine.

Heart disease is the primary contributory factor to the world's rising mortality rate today. There are various risk factors which are responsible for heart diseases. Heart disease risk factors include high blood pressure, overweight, high cholesterol, excessive alcohol use, smoking, an inadequate diet, family history, a lack of physical activity, as well as diabetes. Traditionally, it is difficult to predict heart diseases based on patient's reports which causes a delay in the treatment. Machine learning algorithms work on existing data for prediction of diseases. Machine learning algorithms workremarkably well in predicting heart diseases using fourteen common features. The fourteen features which are chiefly used for heart disease prediction usually involve patient's age, sex, heart rate, chest pain, cholesterol ,blood pressure, fasting blood sugar, electrocardiography findings, slope of the peak exercise ST segment,class value, exercise-induced angina, the number of important vessels coloured by fluoroscopy, value of duration of exercise in minutes, and ST depression induced by exercise. Significant features along with proper machine learning algorithms gives best result in heart disease prediction. Result of prediction also depends upon the size of data. With large amount of data, complexity issues arises. We have done analytical research on previous proposed models. Through our literature survey, we noticed that many researchers while developing their proposed systemmet with the problem of selecting the number of features and proper supervised machine learning algorithm.Some researchers reduced the number of features and achieved high accuracy. We majorly have seen that five prominent features can help to achieve high accuracy with limited amount of data. Random forest gives high accuracy in prediction among all remaining algorithms of our study. In our research paper, we have described the summary of previously proposed model and also discussed the background work done on these four algorithms.

In figure 1 we are explaining how the Machine learning process works. It takes input or old data first and analysed it on the basis of problem statement. Later patterns will be decided on the basis feedback received by using algorithms. In unsupervised machine learning, machine will designed the model on the hidden structures. Following figure 1 includes steps of machine learning process.

Figure 1. Machine learning process

Supervised machine learning algorithms:

Machine learning algorithms broadly categorized into supervised, unsupervised and reinforcement. In our research work we are analysing supervised machine learning algorithms. For analysing algorithms in accuracy point of view we are taking Random forest (RF), Decision Tree (DT), SVM (support vector machine) and KNN (K-nearest neighbour) out of among.

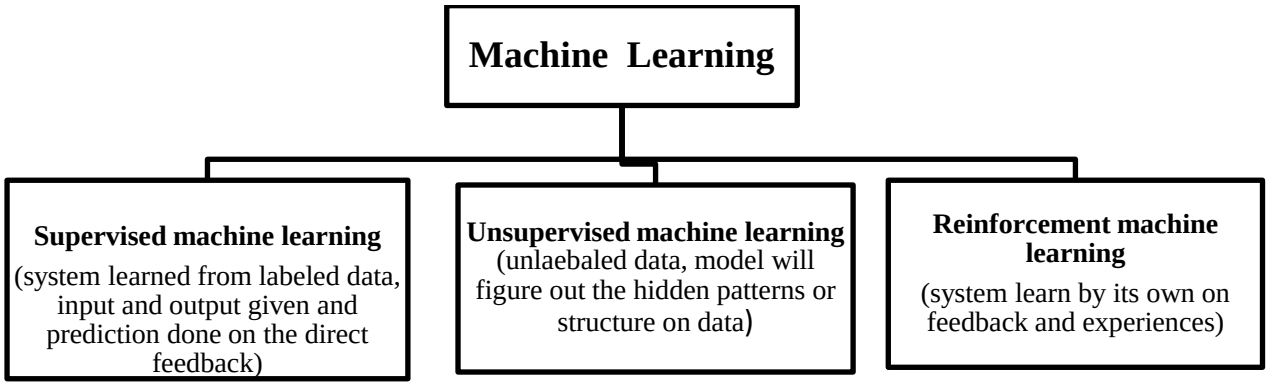

Figure 2. Types of Machine learning algorithms

Random Forest (RF) - A forest is made up of many different types of trees, and the more trees there are, the more robust the forest will be. Similar to this, the accurateness and problem-solving capability of a random forest algorithm improves as the algorithm's number of trees develops. A classifier known as random forest creates several decision trees on various subsets of the input data in order to enrich the dataset's forecasting accuracy. It is based on the idea of ensemble learning, which is the technique of integrating various classifiers to solve a challenging problem and enhance the model's performance. Ensemble is the way of referring to this blending of various models and it can possible through bagging and boosting. Bagging is the process of generating a different training subset via replacement from a sample training dataset. The outcome is decided by a majority vote. Turning weak learners into strong ones by building successive models to achieve maximum possible accuracy is called boosting.

Following are the steps involved in the random forest algorithm

Step 1: Pickrandom instances from the knowndata collection or training set.

Step 2: Build a decision tree for every training set of data.

Step 3: Voting will be conducted using an average of the decision tree.

Step 4: At the end, select the guess result that received the most votes.

In this way random forest works and gives the better accuracy in the prediction of diseases. Another supervised machine learning algorithms is decision tree.

Decision Tree (DT) –One of the supervised machine learning algorithms is decision tree. Decision tree used for classification and regression as well.Decision tree consist with leaf node, root node, decision node and branches. Leaf node shows the probable outcomes. Leaf node does not split into further sub nodes. Whereas decision node having several branches and all are used to generate conclusions.

Decision tree algorithm perform following steps,

Step 1: Start the tree from the root node.

Step 2: Identify dataset's top attribute.

Step 3: Subset the root node to include probableoutcomes as the finest traits.

Step 4: Form the decision tree node which has the finest feature in the step.

Step 5: Use the selections of the dataset generated in step 3 to iteratively develop new decision trees. Continue all the above iterations till you come to that point where you cannot further classify the nodes, and you can denote the last node as a leaf node.

KNN (K-nearest Neighbour) - Agriculture, economics, text mining, and healthcare all make use of KNN algorithm for prediction. The basic goal of the KNN method is to anticipate a new sample point's categorization using data points that have been divided into various groups. It is distance-based and categorises things according to the classes of their close neighbours. KNN is most frequently

utilised for classification tasks. The number of labelled points (neighbours) taken into account for classification is indicated by the parameter k in KNN. The number of these points needed to calculate the outcome is indicated by the value of k. calculating the distance and determining which categories are closest to our unidentified thing are our tasks.

Following are KNN Algorithm steps:

Step 1: Consider labelled data

Labelled data is also called training set which are used to train the model. You can either use the labelled databases found in open sources like this one or manually label data to generate the training set.

Step 2: Searching for K-nearest neighbours

Finding a record's k-nearest neighbours entails locating the records that share the most features with it. This process is often referred to as the distance calculation or similarity search.

Step 3: Group the points

In classification problems, the algorithm chooses a class or group label by a majority vote, which means that the label that appears more frequently in neighbours is used. The K-distance is the separation between a specified query point and a set of data points. We must choose a distance measure in order to calculate it. Distance measurement techniques are several. Hamming distance, Minkowskidistance, Manhattan distance and Euclidean distance are the popular metrics used to measure k-distance.

Support Vector Machine (SVM) -Like decision tree, random forest and K-nearest Neighbour, support vector machine is also used for classification as well as regression.Vladimir Vapnik developed the support vector machine algorithm. In this algorithm, the main objective is to find out "hyperplane". Hyperplane is the accurate plane or line or decision boundary which can classify the number of features (is also refered by n-dimensional space) into sets. In order to build the hyperplane, support vector machine picks the extreme points and vectors. Support vectors are the term used for these extreme examples, and the support vector machine technique is named after them.

Research methods we used:

To incline our research work in the proper direction, we have used specific terms to search the related articles in Google scholar. We used the terms like "heart disease and machine learning algorithm", "healthcare industry and machine learning". For our study we selected research papers which are published after 2016 onwards and in reputed journals like IEEE Access, Science Direct and Elsevier.

Background Research Work:

Javeed A. et al. (2019), in [5] developed the model by hybridizing RSA algorithm with random forest and named RSA-RF. RSA-RF model achieved 93.33% accuracy in heart failure detection and also gained the success to reduce the time complexity. According to the authors, reducing time complexity could be achieved by reducing number features. Authors selected 7 prominent features in RSA-RF.

Human heart sound generated by atrioventricular valve, plays a vital role in detection of any valvular heart disease. According to Tanmay Sinha Roy et al [8] in the year 2022, heart sound is one of the important feature fed for pre-processing of data where noisy signals removed. In [8]authors' opinion, RF is fastest machine learning classifiers among KNN, RF, ANN,SVM and Naïve Byes at run time. Authors further stated that RF has some complexity issues in large amount of data.

In 2019, Senthilkumar et al. [3] combined random forest and linear model together and came with new model. The new model is identified as HRFLM (hybrid random forest with linear model). In [3], authors integrated random forest and linear model for preprocessing of data.According to authors, through HRFLM 88.70% accuracy of predicting heart disease could be achieved without any restriction in features selection [3].

Ensemble is powerful technique which canbe used to increase the accuracyof heart disease prediction [4]. According C.Beulah Chirsta et al.[4] if weak learner can integrated with strong learner then

effectiveness of weak learner can be increased. In[4] ,authors, combined multiple classifiers into proposed system to achieved best performance in prediction of heart disease at early stage.

Identification of proper machine learning techniques with correct selection of remarkable features is very difficult task in prediction [6]. According to Mohammad Shafennor et al. (2018) in [6], researchers are even now facing problems for selecting proper technique with correct set of features. In [6], Authors focused on selecting nine major features with three machine learning algorithms. To find out nine correct features authors used the strategy of analysing, in which they first found how many times the feature get selected in the previous prediction model in respect to achieved accuracy and precision. Secondly authors tried all nine features with combination of seven machine learning algorithms. Finally researchers concluded that with nine major features support vector machine, Naïve Bayes and Vote given highest accuracy in prediction.

Kartik Budholiya et al. in [7], highlighted that optimization of hyper parameters are very essential. Hyper parameters are those parameters, on which learning model will finally determines the values for model and stop determining further learning. Kartik Budholiya et al. in [7], came with their proposed model for predication of heart disease which is based on tuning of hyper parameter and encoding of categorical variables in dataset.Researchers combined Bayesian classification and One-Hot (OH) encoding algorithm. Finding error or missing values in dataset at initial stage will give efficiency in prediction [7].

According to Md.Mamun Ali et al. in 2021[2], 100% accuracy in the prediction of heart disease can be achieved through decision tree, K nearest neighbour and random forest. In [2], in opinion of researchers, chest pain feature is very important in prediction of heart disease.

We are studied and analysed random forest, KNN, DT and SVM algorithms. In some previous research, researchers built their proposed model by integrating with core supervised machine learning algorithms. Through this integration predicting of heart disease as early stage can be achieved with high accuracy rate. Table 1 depicted the summary of machine learning algorithms when they combined with other algorithms to build the proposedmodel as well as when they used independently in prediction of heart disease.

Table 1 The summary of previously used machine learning algorithms

References	Name of the machine learning algorithm used	Combined with other	Proposed model strategy and name	Dataset used for experiments	Performance result of proposed model (% of Accuracy)
[1]	RF,KNN,DT, SVM	-	VLRAKN	UCI repository	83
[2]	RF, DT, KNN,ABM1, LR and MLP	-	-	Cleveland data set from UCI repository	100 (KNN, RF and DT)
[3]	RF,LM	-	Combined RF and LM in HRFLM	UCI repository	88.70
[4]	RF, Bayes Net C4.5, Naïve Bayes, PART and Multilayer perceptron	-	combining Naïve Bayes+ Bayes Net+ RF+ Multilayer Perception	Cleveland data set from UCI repository	85.48
[5]	RF	RSA ,Grid Search	RSA-RF	Cleveland data set from	93.33

				UCI repository	
[6]	KNN, Neural Network, Decision tree, Logistic Regression, Support Vector Model, Naïve Bayes and Vote	-	Combined Vote with Logistic Regression and Naïve Bayes	Cleveland data set and Statlog dataset from UCI repository	87.41
[7]	Bayesian Classification	One-Hot (OH)encoding algorithm	Combined Bayesian classification with OH encoding	Cleveland data set from UCI repository	91.80
[8]	RF, SVM	CNN based deep learning model Inception Net, Residual Net	CNN based xception net model	Samples of Heart sound collected from https://github.com/yaseen21khan/Classification-of-Heart-Sound-Signal-Using-Multiple-Features	99.43

Conclusion:

In the healthcare industry, for the prediction of disease various supervised machine learning algorithms commonly used. The main purpose of our study is to focus on four machine learning algorithms that are widely used. Random forest, support vector machine, decision tree, and K-nearest neighbour are regularly used in the prediction. We found that all these algorithms are flexible and can be combined with other machine learning algorithms to achieve more efficacy. All are finely worked independently as well as when they are hybridized.

To get highest accuracy in prediction of heart disease, selection of flawless features along with best machine learning algorithm is extremely significant. We can use various classifiersof machine learning techniques and also combined with others to get efficient result. We observed that prediction accuracy will be high with minimum number of features selection. We also noticed thatdue to increase in attributes and size of data, time complexity issue arises andthat can be reduced with limited data set and by selecting prominent features for prediction.

Terms which are used in the paper

1) RF – Random Forest
2) SVM- Support Vector Machine
3) DT- Decision Tree
4) KNN- K- nearest neighbour
5) RSA- Random Search Algorithm
6) CNN- Convolutional Neural Network
7) LR- Linear Regression
8) MLP- Multilayer perceptron
9) ABM1- AdaboostM1
10) PART- Projective Adaptive Resonance Theory

References

[1] Chang, V., Bhavani V. R., Xu A. Q., & Hossain M. A. (2022). "An artificial intelligence model for heart disease detection using machine learning algorithms", Healthcare Analytics, 2, 100016.

[2] Ali M. M., Paul B. K., Ahmed K., Bui, F. M. Quinn J. M., & Moni M. A. (2021). "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison". Computers in Biology and Medicine, 136, p.104672.

[3] Mohan S., Thirumalai C., & SrivastavaG. (2019). "Effective heart disease prediction using hybrid machine learning techniques". IEEE access, vol.7, pp.81542-81554.

[4] Latha, C. B. C., & Jeeva S. C. (2019). "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques". Informatics in Medicine Unlocked, vol.16, p.100203.

[5] Javeed A., Zhou S., Yongjian L., Qasim I., Noor A., & Nour R. (2019). "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.

[6] Amin M. S., Chiam Y. K., & Varathan K. D. (2019). "Identification of significant features and data mining techniques in predicting heart disease".Telematics and Informatics, vol.36, pp.82-93.

[7] Budholiya K., Shrivastava S. K., & Sharma V. (2020). "An optimized XGBoost based diagnostic system for effective prediction of heart disease". Journal of King Saud University-Computer and Information Sciences.

[8] Roy T. S., Roy J. K., & Mandal N. (2022). "Classifier identification using Deep Learning and Machine Learning Algorithms for the detection of Valvular Heart diseases". Biomedical Engineering Advances, pp.100035.

[9] Uddin S., Khan A., Hossain M. E., & Moni M. A. (2019). "Comparing different supervised machine learning algorithms for disease prediction". BMC medical informatics and decision making, vol.19 (1), pp.1-16.

[10] Mienye I. D., Sun Y., & Wang Z. (2020). "An improved ensemble learning approach for the prediction of heart disease risk". Informatics in Medicine Unlocked, vol.20, pp.100402.

[11] Saxena K., & Sharma R. (2016). "Efficient heart disease prediction system".Procedia Computer Science, vol.85, pp.962-969.

[12] Tripoliti E. E., Papadopoulos T. G., Karanasiou G. S., Naka K. K., & Fotiadis D. I. (2017). "Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques". Computational and structural biotechnology journal, vol.15, pp.26-47.

[13] Uyar K., & İlhan A. (2017). "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks".Procedia computer science, vol.120, pp.588 -593.

[14] Pathan M. S., Nag A., Pathan M. M., & Dev S. (2022). "Analysing the impact of feature selection on the accuracy of heart disease prediction". Healthcare Analytics, vol.2, pp.100060.

[15] Shailaja K., Seetharamulu B., & Jabbar M. A. (2018, March). "Machine learning in healthcare: A review". In 2018 Second international conference on electronics, communication and aerospace technology (ICECA), pp. 910-914.

[16] Khan M. A. (2020). "An IoT framework for heart disease prediction based on MDCNN classifier". IEEE Access, vol.8, pp.34717-34727.