HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES: A SYSTEMATIC REVIEW

Kiranjit Kaur¹, Munish Saini²

¹,²Department of Computer Engineering and Technology, Guru Nanak Dev University, Amritsar, Punjab, India
Corresponding Author: Kiranjit Kaur
Email: kiranjitkaur166@gmail.com

https://doi.org/10.26782/jmcms.2020.05.00010

Abstract

The key task within the healthcare field is usually the diagnosis of the disease. In case, a disease is actually diagnosed at earlier stage, then many lives might be rescued. Machine learning classification techniques can considerably help the healthcare field just by offering a precise and easy diagnosis of various diseases. Consequently, saving time both formed ical professionals and patients. As heart disease is usually the most recognized killer in the present day, it might be one of the most challenging diseases to diagnose. In this paper, we provide a survey of the various machine learning classification techniques that have been proposed to assist the healthcare professionals in diagnosing the cardiovascular disease. We started by giving the overview of various machine learning techniques along with describing brief definitions of the most commonly used classification techniques to diagnose heart disease. Then, we review representable research works on employing machine learning classification techniques in this field. Furthermore, a detailed comparison table of the surveyed papers is actually presented.

Keywords: Heart Disease, Heart Disease Prediction, Machine Learning, Machine Learning Classification Techniques.

I. Introduction

Nowadays, more people die every year from non-communicable diseases as compared to infectious diseases. The heart related diseases consumes around a million lives of peoples every year, creating this as the primary reason. One death among three is due to heart disease in the United States (US). In the year 2016, around 9,20,000 peoples had heart attacks, and nearly half of them occurred suddenly without prior symptoms. Sudden death is the only symptom for heart disease. Miserably, most of them belonging to young age particularly in India. In India, heart disease happens 1to1.5 decades in advance when compared to the western countries. An estimation reports that there are around 45,000,000 will be affected by heart problems. There seems to be a stable rise in hypertension pervasiveness for the previous 5 decades, which is extra in urban areas than in pastoral zones. It is 25
to 30 percent in urban and 10 to 50 percent in rural zones\cite{I}. Inactive lifestyle is a main reason of death, disease and disability which multiplies the danger of heart disease. In the current days, the heart specialized hospitals executed around 2 Lakhs of surgeries, particularly open hearted every year which is the top most figure worldwide. It increases by 25 to 30 percent every year constantly. Heart disease is the most leading primary reasons of death in the world, demanding around seventeen million people’s lives annually. Heart disease or CVD is a common word to represent the diseases related to the functioning of the heart, arteries, veins and blood vessels. This word is usually taken in diseases like disaster in function of the heart, cardiomyopathy, CAD, exterior vascular ailments, strokes, cardiac arrests and congenital heart diseases. NareshTrehan, the Chairman and the Managing director at Gurgaon’s Medanta told that an upsurge of nearly 10.5% of young patients resides in well-developed cities and 6% percent of them from under developed zones. When comparing similar age group of peoples in the west it is found that around three to four percent increase in our country. So, there is a need for the actual prediction of the heart disease for the quick action to lessen the mortality rate caused due to the various heart problems.

II. Back Ground

This section presents information of the related topics of this paper including machine learning with its techniques along with quick description, data preprocessing, efficiency analysis metrics as well as a concise clarification of the most common cardiovascular disease dataset.

II.i. Machine Learning

Machine learning is used to provide the good learning to the machines and analyze some pattern for handling the data in extra efficient manner. Sometimes, it may happens that after viewing the data, we even unable to predict the actual pattern or acquire the valuable information from the data. In this condition, we have to go for machine learning\cite{XXVI}. The motive of machine learning is to grasp some knowledge from the data by themselves. Even, many studies has been terminated which highlights the purpose of machine learning that how do machines learn by its own\cite{XXXIII},\cite{III}.

All the algorithms with their representation have been explained in the upcoming content of this paper.

II.ii. Machine Learning Techniques

The main ML techniques can be classified as follows:

II.ii.a. Supervised Learning

The supervised machine learning algorithms are those which demand some external assistance. The input dataset splits into training and test dataset. The trained dataset composed of output variable which is to be predicted or classified. Each algorithm get to know a specific pattern from the training dataset and just apply them to the test dataset for prediction or classification purposes\cite{XVIII}. This algorithm is named as supervised learning in view of the fact that the process of an algorithm
learning from the training dataset can be thought of as a teacher supervising the learning process. Three most prominent supervised learning algorithms are considered below.

1) **Decision Tree:** Decision tree is the type of tree which usually groups attributes simply by sorting them dependent on their particular values. Mainly, Decision tree is required for some classification purposes. Every tree contains a number of nodes and branches. Each node personifies an attribute in a group which is to be classified along with which each branch symbolizes a value that the node can take. An illustration for decision tree is provided in Figure 1.

![Figure 1: Decision Tree](XXVI)

2) **Naïve Bayes:** Naïve Bayes is an uncomplicated but surprisingly strong algorithm for anticipating the models. It is the one which is mainly focusing on the clustering and classification purposes. The representation for Naïve Bayes is probabilities. A list of probabilities are to be stored to a file for a learned Naïve Bayes model.

This includes:

- **Class Probabilities:** It refers to the probabilities of each class in the training dataset.
- **Conditional Probabilities:** The conditional probabilities of each input value gives each class value. An example of the Bayesian network is given in Figure 2.

![Figure 2: An Example of Bayesian Network](XIV)
3) **Support Vector Machine:** Support vector machine is a set of associated supervised learning methods used for classification and regression. They mainly belong to a family of generalized linear classifiers. In other words, SVM is a classification and regression prediction tool that makes use of machine learning theory to increase the predictive accuracy while automatically avoiding over-fitting to the data. SVM’s were mainly developed for solving the classification problem, but recently it has been extended to solve regression problems. The functioning of SVM is provided below in Fig.3.

![Fig.3: Functioning of SVM(Support Vector Machine)[XXII]](image)

**II.ii.b. Unsupervised Learning**

The unsupervised learning algorithm learns few features from the data. When some new data is organized, it makes use of the previously learned features to identify the class of the data. It is typically used for clustering and feature reduction. The workflow of unsupervised learning is prescribed in figure 4.

![Fig.4: Workflow of Unsupervised Learning[XV]](image)

The two foremost algorithms for dimensionality reduction techniques and clustering are discussed below.
1) **K-Means Clustering:** Clustering is a class of unsupervised learning technique which generates groups automatically during its initiation. The modules which are having homogeneous characteristics are categorized in the same cluster. Ask distinct clusters are being created that is why the algorithm is called k-means clustering[XXVIII]. A clustered data is shown in the figure 5.

![Fig.5: K-means Clustering[XIII]](image)

2. **Principal Component Analysis**

In PCA, the dimensions of the data are minimized just to make the computation quicker and simple. This technique is mainly used for feature selection that is selecting the relevant features from a large dataset which contains a number of attributes. To recognize how PCA works, assume any 2D data.

When the assumed data is being plotted on a graph, it will take up two axes. But eventually, when PCA is put to use on that 2D data, the data as a result will be 1D, which is well explained in the given Figure 6.

![Fig.6: Result after applying PCA on Iris dataset[XII]](image)

II.ii.c. **Semi –Supervised Learning**

Semi-Supervised learning algorithms is an approach that is a combination of both supervised and unsupervised learning. It is much helpful in the domain of data mining and machine learning and where the unlabeled data is previously existing and getting the labeled data is one of the most monotonous process[II][XXXIV]. Semi-supervised learning falls into some important categories[XXXV] which includes:
1) Generative Models
2) Self-Training
3) Transductive SVM

II.i.d. Reinforcement Learning

Reinforcement learning is a class of learning that generate some settlement based on the fact that which actions are to be taken such that the outgrowth will be more favorable. The learner is unaware about which actions to be taken until it’s been provided with a situation. The action that is taken by the learner may affect the situations and their actions in the future. Reinforcement learning solely depends on two criteria: trial and error searching and putting off an outcome [XXX]. The general model [XVI] for reinforcement learning is shown in Figure 7.

![Reinforcement Learning Model](image)

Fig. 7: Reinforcement Learning Model [XVI]

In the above figure, the agent collect an input ‘r,’ and current state ‘s,’ from the environment. Based on these inputs, the agent propagate a behavior and takes an immediate action ‘a,’ which gives outcome.

II.i.e. Ensemble Learning

When a number of individual learners are combined to form only a single learner then that particular type of learning is termed as ensemble learning. It may includes decision tree, Naïve Bayes, neural network and so on. It is a seductive topic from many years. So, It has been observed that the particular task performed by a collection of learners is more reliable rather than taking into account by an individual learner [XVI].

Two most popular Ensembling techniques are mentioned below [XVI]:

1) **Boosting:** Boosting is one of the technique that aims to lessen the bias and variance. Boosting creates a collection of the weak learners along with which it converts these to one strong learner. A weak learner is often a classifier that is hardly correlated to that of true classification and strong learner is highly correlated [IX].
2) **Bagging**: Bagging or bootstrap aggregating is mainly put into practice when we need to increment the veracity and stability of a machine learning algorithm. It is applicable in classification and regression. Bagging also helps in decreasing the variance value and even useful in handling overfitting[X].

II.ii.f. **Multitask Learning**

Multitask learning has a manageable aim of just serving other learners to execute in more effective way. When learning algorithm multitasks is exercised on a certain task, it just recollects the procedure how it finds a solution to the given problem or how it gets to a particular inference. Then this course of action is to be taken by algorithm to acquire the relevant solution of other identical problems. If the learners dispense its progressive experience just with one another, it in return also provides good knowledge to its own self. So, it has been concluded that the learners can analyze good things and learn better when they work concurrently rather than following the individual pattern[IV].

II.ii.g. **Neural Network Learning**

The neural network is derived from the biological concept of Neurons. A Neuron is a cell-like structure in the brain. A Neuron has mainly four parts (Figure 8) which includes dendrites, nucleus, soma and axon.

![Fig.8: A Neuron](image)

The dendrites receive the electrical signals which are sent to the Soma for processing them. The output produced after processing the electrical signals is dispatched by the axon to the dendrite terminals. Additionally, the obtained output is sent on to next neuron. The nucleus present in the center is the heart of the neuron. The interconnection is well named as neural network.

![Fig.9: Structure of an Artificial Neural Network](image)
An ANN works on three layers as mentioned in the fig.11. The input layer which accepts the input (like dendrites). The hidden layer works on processing the input (much like soma and axon). Finally, the output layer which forwards the calculated output (as dendrite terminals)[XXIX]. Additionally, three categories fall under ANN are discussed below[VIII].

1) **Supervised Neural Network:** In this, we already have the output of the input. The predicted output of the neural network is being differentiated with the actual known output. After concluding the error, the parameters are changed, and then again it is fed into the neural network.

2) **Unsupervised Neural Network:** In Unsupervised Neural Network, there is no preceding clue about the output of the input. The main contribution of the network is just to designate the data according to some factual similarities.

3) **Reinforced Neural Network:** In RNN, the neural network behaves same as a human communicates with the environment. From the environment, some behavioral response is provided to the network acknowledging the fact that whether the decision undertaken by the network is exact or not. RNN is represented in Fig. 10.

![Fig.10: Reinforced Neural Network][VIII]

II.ii.h. **Instance-based Learning**

In instance-based learning, the learner tries to learn some pattern from the existing data. After analyzing the pattern from the prevailing data, it attempts to bear on the same pattern to the newly fed dataset. The complexity of the learning algorithm is directly proportional to the size of the data that is the complexity escalates as the size of the data increases. A k-nearest neighbor is a well-known example that is described below[XI].

1) **K-Nearest Neighbor:** In KNN, the training data (which is well-labeled) is fed into the learner. When the test data is introduced to the learner, it tries to compare both the data. k-most correlated data is to be taken from the training set. The majority of k is taken which serves as the new class for the test data[VII].
II.iii. Data Preprocessing

The actual overall performance in addition to accuracy of the predictive model is not just impacted by the actual algorithms applied, but additionally by the expertise of the dataset along with the preprocessing techniques. Preprocessing signifies all the steps applied on the dataset before using any machine learning algorithm to the provided dataset. The preprocessing phase is extremely important mainly because it works on the dataset and applies the idea in a form the fact that algorithm understands.

Datasets might have faults, missing out on information, redundancies, noise, and several other concerns which usually trigger the data for being unsuitable to be used by the machine learning algorithm directly. An additional aspect is actually the size of the dataset. Quite a few datasets have numerous attributes making it more difficult for the algorithm to examine it, locate patterns, or even generate precise predictions. This sort of difficulties might be sorted out by analyzing the dataset and making use of the appropriate data preprocessing techniques. Data preprocessing steps involves: data cleaning, data transformation, missing values imputation, data normalization, feature selection, as well as other steps with respect to the nature of the dataset.

II.iv. Performance Evaluation Metrics

The following are the metrics by which various researcher evaluate the prediction models and describe the performance of their results. We provide a short definition for each method without delving into the deep details and mathematical equations.

1) Accuracy: This metric shows the percentage of the accurate results.
2) Precision: This metric shows how relevant the result is.
3) Recall or Sensitivity: Measures the returned relevant results.
4) F-Measure: Combining precision and recall.
5) Coefficient of determination.
6) Testing time: Total time taken for testing.
7) Root Mean Square Error.
III. Literature Review

Otoom et al.[XXXI] offered a system intended for evaluation and follow-up. Heart disease is actually diagnosed and examined from the proposed system. Dataset was taken from the UCI repository of Cleveland. This dataset contains 303 cases along with 76 attributes. 13 attributes were used out of 76 attributes. A pair of tests by using about three algorithm named Bayes Naive, Support vector machine, and Functional Tree algorithm was carried out for detection purposes. The WEKA tool was used for the same mentioned above. About 83.8% was obtained after the holdout test by making use of SVM technique. Finally, after applying the above mentioned test to the best selected 7 features, SVM achieved the highest accuracy followed by Naive Bayes along with FT having accuracy of 85.1%, 84.5% and 84.5% respectively.

Parthiban et al.[XXIV] diagnosed cardiovascular disease within the patients suffering from diabetes by making use of automatic learning methods by gathering dataset of 500 patients from Chennai Research Institute. Naïve Bayes and SVM algorithms was applied by using WEKA tool. The Naïve Bayes algorithm gives accuracy of 74% and SVM offers the highest accuracy with 94.60%.

Chaurasia et al.[V] recommended to make use of data minings strategies to identify cardiovascular disease. The tool named WEKA was employed that contains some machine learning algorithms intended for mining purposes. The algorithms used for the heart disease prediction includes Naive Bayes, J48 and bagging. The dataset was collected from UCI Repository contains 76 attributes out of which only 11 attributes were taken into concern. As a result, bagging with 85.03% gives a highest accuracy while J48 offers 84.35% accuracy and at last Naive Bayes provides 82.31% accuracy.

Vembandasamy et al.[XXXII] diagnosed the heart disease prediction by making use of the Naïve Bayes algorithm. The dataset was taken from the Chennai institute which contains the record of 500 patients. WEKA tool was taken into account for the prediction. As a result, Naive Bayes gives 86.419% accuracy.

X. Liu et al.[XX] provided a research to help in the detection of cardiovascular disease by making use of a hybrid classification system depending on the Relief and Rough Set method. The above system involved two sub-system named as the actual RFRS system for feature selection and a classification system along with a general classifier. With the cross validation technique of jackknife, 92.59% accuracy was achieved.

A. Malav et al.[XXI] offers a highly effective hybrid algorithmic approach for the prediction of cardiovascular disease, as a way to identify and figure out unidentified knowledge about heart disease by making use of hybrid approach in which both artificial network works with the Naive Bayes. As a result, the accuracy achieved was 97%.

Chen, A.H et al.[VI] proposed a system for the prediction of the heart disease named heart disease prediction system(HDPS) just to predict the heart disease in amore accurate and efficient way. A predictive model was employed to identify the disease together with data and knowledge. Two main approaches utilized for the
The proposed work includes Statistics and machine learning. The actual algorithm included has three techniques: data selection, ANN and Afterward, to trained the data, the Learning Vector Quantization (LVQ) was applied. ROC curve was used for examining the precision of results. Eventually, the actual accuracy obtained by the above-mentioned techniques was 80%.

Jabbar et al. [XXVII] employed the associative classification algorithm intended for the prediction of the cardiovascular disease. The genetic approach was taken into account as the actual algorithm for prediction. Initially, an associative classification was utilized for the classification of the dataset with labeled classes along with which some rules were collected from the training dataset.

Table 1: Comparison of ML classification techniques for heart disease prediction

| Author               | Dataset                                | Tool       | Classification Technique used              | Best Technique found | Accuracy Achieved |
|----------------------|----------------------------------------|------------|--------------------------------------------|----------------------|-------------------|
| Otoom et al. [XXXI]  | Cleveland (UCI) 303 cases, 76 attributes | WEKA       | Naïve Bayes, Support vector machine, and FT | SVM                  | 88.3%             |
| Parthiban et al. [XXIV] | Chennai Research Institute (500 patients data) | WEKA       | Naïve Bayes, SVM                           | SVM                  | 94.60%            |
| Chaurasia et al. [V] | UCI machine learning laboratory       | WEKA       | Naïve Bayes, J48 and Bagging               | Bagging              | 85.03%            |
| Vembandasamy et al. [XXXII] | Chennai Research Institute (500 patients data) | WEKA       | Naïve Bayes                               | Naïve Bayes          | 86.419%           |
| X. Liu et al. [XX]   | UCI machine learning laboratory       | Not Mentioned | Relief and Rough set (RFRS) method        | cross-validation scheme | 92.59%           |
IV. Gaps in Literature

From the existing literature, it has been found that existing machine learning models suffer from at least one of the following issues:

1) **Feature Selection:** Majority of existing researchers have neglected the effect of feature selection techniques during the training and testing time. It has been observed from the literature that an efficient feature selection technique has an ability to enhance the performance of machine learning models.

2) **Ensembling:** Most existing researchers have neglected the use of optimistic ensembling approaches to enhance the performance of existing machine learning models for Heart disease prediction.

3) **Parameters Tuning:** Parameters tuning is another major gap found in the existing literature. An efficient tuning of parameters has ability to improve the performance further.

4) **Meta-Heuristic Techniques:** It has been observed that majority of existing researchers have focused on designing heuristic machine learning model to predict antibody Heart disease prediction.

V. Conclusion

This paper describes the literature of various machine learning techniques for the prediction of heart disease. The accuracy of the proposed models may vary and it depends on the quality of dataset used, tool used by various researchers, the number of attributes and records in the dataset along with the preprocessing techniques used in the model. It depends on whether it is a hybrid model or not and whether the model make use of feature selection or not. From comparison table, we can conclude that the researcher who produced the highest accuracy was Malav that uses Hybrid approach with combining K-means clustering algorithm and ANN by making use of WEKA tool and dataset was taken from Cleveland UCI repository. The dataset must be

---

[n/a: not applicable]
preprocessed for getting good results. Also, a suitable algorithm must be used when developing a prediction model.

Finally, machine learning used for diagnosing heart disease which helps both healthcare professionals and patients. It is still working for various fields. As observed from the comparison table, most of the researchers got the same dataset from the same source which is the UCI repository. So, there is a requirement for more high-quality datasets that will be published by various hospitals so that researchers can have a good source for their prediction for the various diseases which helps in obtaining the good results with high accuracy.

References

I. Alsabti, K., Ranka, S., & Singh, V. (1997). An efficient k-means clustering algorithm.

II. Andrecut, M. (2009). Parallel GPU implementation of iterative PCA algorithms. *Journal of Computational Biology, 16*(11), 1593-1599.

III. Bowles, M. (2015). Machine learning in Python: essential techniques for predictive analysis. John Wiley & Sons.

IV. Caruana, R. (1997). Multitask learning. *Machine learning, 28*(1), 41-75.

V. Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. *International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2*, 56-66.

VI. Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011, September). HDPS: Heart disease prediction system. In *2011 Computing in Cardiology* (pp. 557-560). IEEE.

VII. Harrington, P. (2012). *Machine learning in action*. Manning Publications Co..

VIII. Hiregoudar, S. B., Manjunath, K., & Patil, K. S. (2014). A survey: research summary on neural networks. *International Journal of Research in Engineering and Technology, 3*(15), 385-389.

IX. https://en.wikipedia.org/wiki/Boosting_(machine_learning)

X. https://en.wikipedia.org/wiki/Bootstrap_aggregating

XI. https://en.wikipedia.org/wiki/Instance-based_learning

XII. https://en.wikipedia.org/wiki/Principal_component_analysis

XIII. http://pypr.sourceforge.net/kmeans.html

XIV. https://webdocs.cs.ualberta.ca/~rgreiner/C-651/Homework2_Fall2008.html

XV. http://www.simplilearn.com/what-is-machine-learning-and-why-it-matters-article
Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research, 4*, 237-285.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics, (4)*, 580-585.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160, 3-24.

Lowd, D., & Domingos, P. (2005, August). Naive Bayes models for probability estimation. In *Proceedings of the 22nd international conference on Machine learning* (pp. 529-536).

Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. (2017). A hybrid classification system for heart disease diagnosis based on the RFRS method. *Computational and mathematical methods in medicine, 2017*.

Malav, A., Kadam, K., & Kamat, P. (2017). Prediction of heart disease using k-means and artificial neural network as Hybrid Approach to Improve Accuracy. *International Journal of Engineering and Technology, 9*(4), 3081-3085.

Meyer, D., & Wien, F. T. (2015). Support vector machines. *The Interface to libsvm in package e1071*, 28.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research, 11*, 169-198.[21] Zhou, Z. H. (2009). Ensemble Learning. *Encyclopedia of biometrics, 1*, 270-273.

Parthiban, G., & Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. *International Journal of Applied Information Systems (IJIAS)*, 3(7).

Richert, W. (2013). Building machine learning systems with Python. Packt Publishing Ltd.

Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4), 476-487.

Salem, T. (2018). Study and analysis of prediction model for heart disease: an optimization approach using genetic algorithm. *International Journal of Pure and Applied Mathematics, 119*(16), 5323-5336.

Shalev-Shwartz, S., Singer, Y., & Srebro, N. Pegasos: Primal estimated subgradient solver for svm 2007b. URL http://ttic.uchicago.edu/shai/papers/ShalevSiSr07.pdf. A fast online algorithm for solving the linear svm in primal using sub-gradients.
XXIX. Sharma, V., Rai, S., &Dev, A. (2012). A comprehensive study of artificial neural networks. *International Journal of Advanced research in computer science and software engineering*, 2(10).

XXX. Sutton, R. S. (1992). Introduction: The challenge of reinforcement learning. In *Reinforcement Learning* (pp. 1-3). Springer, Boston, MA.

XXXI. Tan, K. C., Teoh, E. J., Yu, Q., &Goh, K. C. (2009). A hybrid evolutionary algorithm for attribute selection in data mining. *Expert Systems with Applications*, 36(4), 8616-8630.

XXXII. Vembandasamy, K., Sasipriya, R., &Deepa, E. (2015). Heart diseases detection using Naive Bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, 2(9), 441-444.

XXXIII. Welling, M. (2011). A first encounter with Machine Learning. *Irvine, CA.: University of California*, 12.

XXXIV. Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1), 1-130.

XXXV. Zhu, X. J. (2005). *Semi-supervised learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.