Soft Computing, Data Mining, and Machine Learning approaches in Detection of Heart Disease: A Review

Keshav Srivastava[a] and Dilip Kumar Choubey[b]

[a,b]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore
[a]{keshav.srivastava1981@gmail.com,
keshav.srivastava2019@vitstudent.ac.in}
[b]{dilipchoubey_1988@yahoo.in, dilip.choubey@vit.ac.in}

Abstract. Heart disease detection is the need of the hour as it not only deteriorated adults but children are also showing symptoms of it all over the world. It can occur to a person having an improper diet, high cholesterol level, smoking habits, addiction to alcohol or drugs and even occurs to a diabetic patient. Various approaches are there in various fields, say in Machine Learning, Soft Computing, Data Mining are there. This paper aims to provide a survey of several research papers comprising of the above techniques on determining the heart diseases. This paper gives the perspective for the researchers for future work.

Keywords: Heart Disease, Machine Learning, Data Mining, Soft Computing, Decision Trees, KNN, Random Forest, Genetic Algorithm, Neural Network, Multilayer Perceptron, SVM, Naïve Bayes Classifier, Classification, Logistic Regression, Backpropagation Algorithm.

1. Introduction

Heart disease can be present since birth or can occur at any age if an individual neglects his appetite or starts consuming alcohol, drugs or smoking. These are of various types like Congenital heart disease: occurs in an individual since birth, Arrhythmia: Glitching in Heartbeat, Coronary Artery Disease: occurs due to deficiency in nutrients and oxygen in blood, Dilated Cardiomyopathy: results in weakening of muscles which results in improper supply of blood in the body. However, the above mentioned can be prevented by having a balanced diet, exercising regularly, maintaining a good BMI, quit smoking, reducing alcohol consumption, controlling high blood pressure and diabetes. A major challenge for the hospitals is how to provide the best treatment to the patient at an affordable cost for which computer based information can be used by making the use of machine learning, soft computing, and data mining. Here we have reviewed various research papers majorly from the last 3 years which provides a solution of detecting heart disease at an early stage and some models can do it without taking the help of any professional i.e. the patients can themselves set up the environment for this. In the same way Choubey et al. (2016, 2017, 2018, 2019) ([3], [7], [10], [14], [19]), Bala et al. (2017, 2018) ([22], [27]) have briefly summarized of many soft computing, data mining, machine learning approaches for classification of Diabetes, and Thunderstorm respectively. Our work will give future researchers a brief idea of what current technologies have been used for the detection of heart disease and what more can be done to improve them. This paper has been organized as follows: Literature Reviews are presented in Section 2 which includes dataset, techniques, tools, advantages, issues, and accuracy. Discussion and future directions are committed to section 3 in which the existing work and future work is presented and thereby concluded the future directions of this study.

2. Literature Review

Here we have studied several research papers that used soft computing, machine learning, and data mining methods. The study of this work emphasis on particularly heart disease classification.

The existing works have been summarized in the form of table which consists dataset used, techniques used, tool used, advantages, issues, and accuracy.

Table 1. Summary of Existing Work for Heart Disease

Paper Reference No.	Dataset Used	Techniques Used	Tool Used	Advantages	Issues	Accuracy
[1]	Hospitals from Andhra Pradesh, India.	KNN, Genetic Algorithm.	--	Improved accuracy.	Not performed well for Breast cancer and primary tumour.	90.7%
[2]	Cleveland Heart Disease dataset.	Random Forest, Big Data, Spark.	Apache Spark	Using a wearable device, real-time data of an individual is processed for mapping with the dataset.	As data was stored in a distributed way so a failure in one system will result in the entire system to fail.	87.5%
[4]	Cleveland Heart Disease Dataset.	Multilayer Perceptron.	PyCharm IDE, Python.	Effective way of determining disease as compared to the expensive ECG, CT scan.	If the Fitbit gives wrong then the result will lead to unnecessary treatment.	--
[5]	Cleveland Heart Disease Dataset.	IoT, Naïve Bayes, K-Nearest Neighbours, Decision Tree, SVM.	Arduino IDE	Real-time display of ECG in an iPhone.	The System is not heterogeneous.	82.90%
[6]	American Heart Association.	Data Mining, Genetic Algorithm.	MATLAB R2012a	System itself halts the training after reaching a minimum error.	Cannot determine the disease in an early stage.	96.2%
[8]	Cleveland Heart	Data Mining, Decision Trees, Naïve	.NET	Using Decision Tree and Naïve Bayes Classifier	It uses categorical data	Neural Network gives 49.34%, Naïve Bayes

	Disease database.	Bayes, and Neural Network.		the most influencing factor is evaluated which is chest pain type.	and the dataset needs to be expanded.	gives 47.58% and Decision Tree gives 41.85%.
[9]	Real-time data from Smart Mouse, Smart Chair, Smart Mirror.	Classification, Neural Network.	Arduino IDE	Simple assets are used to collect data like Mouse, Chair, Mirror and a Smartphone for displaying output.	Data can be inconsistent as it is taken from a variety of sources.	--
[11]	Cleveland heart dataset.	Random Forest, Naïve Bayes, C4.5, Multilayer Perceptron.	--	Using ensemble classification for weak classifier, a 7% increase in accuracy is observed.	Necessary measures are taken to avoid overfitting because of random forest.	76.57% with C4.5, 82.18% with Random Forest, 80.86% with MLP, 84.49% with Naïve Bayes.
[12]	Cleveland Clinical Foundation Heart Disease dataset.	Decision Tree, Logistic Regression, Random Forest.	Python IDE	For better accuracy, Hyper parameter Tuning is applied.	By taking default values of hyper parameter, decision tree model over fits the data.	Decision Tree gives 92.59%, Logistic Regression gives 88.50%, and Random Forest gives 93.61%.
[13]	Italian Dataset and American Dataset by NIDDK Repository.	Non-Linear SVC, RBF Kernel Algorithm, Grid Search Algorithm.	--	Missing Values are rectified.	Attributes occurring in both dataset are taken.	95.25% in Italian Dataset and 92.15% in American Dataset.
[15]	Cleveland Heart Disease	Genetic algorithm, K-means algorithm,	--	MAFIA algorithm helped in handling huge dataset.	No alarming system is there.	--

	Database.	MAFIA algorithm, Decision tree classification.				
[16]	Symptoms are collected by Doctor.	Data Mining, K means clustering, SVM Algorithm, Decision Tree.	--	SVM algorithm can find the relationship between the variables of the dataset.	SVM generates Quadratic Programming problem.	--
[17]	Heart Disease Data Warehouse.	Naïve Bayes Classifier, KNN, 10-fold cross validation.	Tangara	Tangara improves the accuracy of the model.	Data comes from various sources so it may be possible that the training data is exploited.	Naïve Bayes gives 52.33%, Decision List gives 52%, and KNN gives 45.67%.
[18]	Cleveland heart disease dataset.	Logistic Regression, SVM, Naïve Bayes, ANN, Decision Tree, KNN, Random Forest.	--	Various feature selection algorithms are used to find out the irreverent features.	6 samples are removed in dataset because of missing values.	Logistic Regression gives 84%, SVM gives 86%, Naïve Bayes gives 83%, ANN gives 74%, Decision Tree gives 74%, KNN gives 76%, and Random Forest gives 83%.
[20]	Cleveland Heart Disease database & Statlog Heart Disease database.	Data Mining, Decision Trees, Naïve Bayes, Neural Network.	Weka 3.6.6	Data pre-processing is used.	To mine unstructured data, text mining can be used.	Decision Trees gives 96.66%, Naïve Bayes gives 94.44%, and Neural Network gives 99.25%.

[21]	Real-time data is taken by the wearable device.	Machine Learning, Java.	--	Negative alerts aware of the patient and the usage of medication prescription feature.	For effective result, patient has to wear the device 24X7.	--
[23]	Hungarian data, Cleveland data, and Switzerland data.	Soft Computing, Data Mining, Neural Network and Machine Learning.	MATLAB	Model predicts the disease using more than one dataset	Varying value of 'k' is chosen for all the datasets.	57.85%
[24]	Pima Indians diabetes and Cleveland Heart Disease.	Fuzzy Logic, Data Mining, Backpropagation Algorithm, K fold cross validation.	--	Local error is estimated to improve the classification.	For diagnosing new patients extracted rules are used.	Pima Indians diabetes gives 84.2% and Cleveland Heart Disease gives 86.8%.
[25]	Ulster Hospital of Northern Ireland.	Genetic Algorithm, Weighted KNN.	--	Weighted KNN is used to boost the effectiveness.	A larger dataset could have been used to explore the model.	63.56%
[26]	Cleveland Heart-Disease Database, Cardiology inpatient dataset from PKU People's Hospital.	Random Forest, Classification, Ensemble.	--	Cleaning, Integration and Standardization is done before professing the data.	In Cleveland dataset, there are many redundant values which can be rectified using Data Mining.	Cleveland Heart-Disease Database gives 91.6%, Cardiology inpatient dataset gives 97%.

3. Discussion and Future Direction

Here are some papers which are reviewed on the prediction and detection of Various Heart Disease like Congenital heart disease, Arrhythmia, Coronary Artery Disease[25], Dilated Cardiomyopathy using various algorithms of neural network[6], machine learning ([4], [5], [12], [13], [25]) and some data mining techniques ([6], [8], [15], [20]). In some papers a hybrid system[24] is made to detect the disease, some models offer the real-time heart rate, ECG, some models use IoT technologies. The Table 2 illustrating the future work over existing work for further implementation and research purposes.

Table 2. Summary of Future Work over Existing Work

Paper Reference No.	Existing Work	Future Work
[4]	A fit-bit is used to determine the heart disease in real-time using multi perceptron model which is an economical as compared to expensive ECG and CT Scan.	Using Cloud Technologies the data can be stored on the cloud and thus can be used to determine like cancer, diabetes using machine learning, image processing, and fuzzy logic techniques.
[5]	By combining the power of IoT and machine learning algorithms like Naïve Bayes, K-Nearest Neighbours, Decision Tree, and SVM a real-time ECG is obtained on the user's phone.	As the data is taken in real-time so deep learning can be used to make the data more effective by updating weights and calculating error to enhance the accuracy.
[6]	This model utilizes the true power of machine learning algorithms as usually a dataset contains a lot of irrelevant attributes but using Logistic Regression, SVM, Naïve Bayes, ANN, Decision Tree, KNN, Random Forest the irreverent features are removed.	A k cross validation algorithm can be implemented on the dataset to compensate for the missing values.
[8]	A huge dataset is taken on which Decision Tress and Naïve Bayes Classifier are applied and it is concluded that chest pain type is most important in determining the heart disease.	Other medical attributes can be used along with other data mining techniques like Clustering, Association Rules and Time Series on continuous dataset instead of categorical dataset for better performance.
[12]	Various Machine Learning algorithms were applied on the datasets of which Random Forest gives the best result.	Deep Learning can be applied to the system to find effective errors and thus boost the effectiveness of the system.

[13]	Algorithms like Non-Linear SVC, RBF Kernel Algorithm, and Grid Search Algorithm were applied on both the datasets to remove the missing values.	Various Data Mining and Soft Computing techniques can be applied to the Italian Dataset to fully exclude the bias value.
[15]	A real-time heart rate monitoring system is developed and the patient is going to have a heart attack then an emergency message will be sent.	A health report summary can be generated for future work with which the patient can go to the doctor for better consultation and the number of recipient for an emergency should not be limited to ambulance only.
[20]	This model uses the Data Mining, Decision Trees, Naïve Bayes, Neural Network using which Data pre-processing is done on the dataset to find out the missing values.	Data Mining is not enough to extract features from the dataset as this model can use Text mining to mine the unstructured data present in the dataset.
[24]	A hybrid system is developed using Fuzzy Logic, Data Mining, Backpropagation Algorithm, K fold cross validation for the diagnosing of heart disease in which classification is improved by estimating local error.	For treating new patients some knowledge can be extracted from the trained hybrid neural network.
[25]	Data from Ulster Hospital of Northern Ireland is obtained and Genetic Algorithm, Weighted KNN are used to find the heart disease of which around 20% of the missing values are obtained using Weighted KNN.	For future work, the implicit or explicit knowledge can be used for the comparison of the different models and the data must be pre-processed for incorporating with the missing values.

4. Conclusion

As the number of persons suffering from heart disease are increasing due to various factors like diet, heredity, smoking habits, cholesterol and diabetes so it is a dire need to make an effective system which can determine any heart disease at a really early stage so that the proper treatment can be done and the price of developing that system should also be taken into consideration so that a person belonging to any background can utilize it. As discussed in this paper majorly Machine Learning, Soft Computing and Data Mining are taken into consideration which provided useful information and some model attained an accuracy greater than 90%. From this study one of the future direction, may be to design such a classification based model which will provide more accuracy then existing as shown in Table 1. Future researchers can also use Big Data technologies along with these to make the model more efficient and reliable to users as well as developers.

References

[1] Jabbar, M.Akhil., Deekshatulu, B.L., Chandra, Priti. (2013). Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. Elsevier, pp: 85-94.

[2] Ed-daoudy, Abderrahmane., Maalmi, Khalil. (2019). Real-time machine learning for early detection of heart disease using big data approach. IEEE.

[3] Choubey, Dilip Kumar., Tripathi, Sudhakar., Kumar, Prabhat., Shukla, Vaibhav., Dhandhania, Vinay Kumar. Classification of Diabetes by Kernel based SVM with PSO, Recent Patents on Computer Science, Bentham Science, Accepted (In Press, 2019).

[4] Gavhane, Aditi., Kokkula, Gouthami., Pandya, Isha., Devadkar, Prof. Kailas. (2018). Prediction of Heart Disease Using Machine Learning. IEEE, pp: 1275-1275.

[5] Dharmasiri, N.D.K.G., Vasanthapriyan, S. (2018). Approach to Heart Diseases Diagnosis and Monitoring through Machine Learning and iOS Mobile Application. IEEE, pp: 407-412.

[6] Amin, Syed Umar., Agarwal, Kavita., Beg, Dr. Rizwan. (2013). Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors. IEEE, pp: 1227-1331.

[7] Choubey, Dilip Kumar., Paul, Sanchita., Sandilya, Smita., Dhandhania, Vinay Kumar. Implementation and Analysis of Classification algorithms for Diabetes, Current Medical Imaging Reviews, Bentham Science, Accepted (In Press, 2018).

[8] Palaniappan, Sellappan., Awang, Rafiah. (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. IEEE, pp: 108-115.

[9] Wijaya, Rifki., Prihatmanto, Ary Setijadi., Kuspriyanto. (2013). Preliminary Design of Estimation Heart disease by using machine learning ANN within one year. IEEE.

[10] Choubey, Dilip Kumar., Paul, Sanchita. (2017). GA_RBF NN: A Classificatio System for Diabetes. International Journal of Biomedical Engineering and Technology (IJBET), Inderscience, Vol. 23, No. 1, pp. 71-93.

[11] Latha, C. Beulah Christalin., Jeeva, S. Carolin. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Elsevier.

[12] R, Shashikant., P, Chetankumar. (2019). Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter. Elsevier.

[13] Mezzatesta, Sabrina., Torino, Claudia., Meo, Pasquale De., Fiumara, Giacomo., Vilasi, Antonio. (2019). A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. Elsevier.

[14] Choubey, Dilip Kumar., Paul, Sanchita. (2016). Classification techniques for diagnosis of diabetes: a review, International Journal of Biomedical Engineering and Technology (IJBET), Inderscience, Vol. 21, No. 1, pp. 15-39.

[15] Babu, Sarath., EM, Vivek., KP, Famina., K, Fida., P, Aswathi., M, Shanid., M, Hena. (2017). Heart Disease Diagnosis Using Data Mining Technique. IEEE, pp: 750-753.

[16] Raju, Cincy., E, Philipsy., Chacko, Siji., Suresh, L Padma., S, Deepa Rajan. (2018). A Survey on Predicting Heart Disease using Data Mining Techniques.

IEEE, pp: 253-255.
[17] Rajkumar, Asha., Reena, Mrs. G.Sophia. (2010). Diagnosis Of Heart Disease Using Datamining Algorithm. GJCST, Volume: 10, Issue: 10, Version: 1.0, pp: 38-43.
[18] Haq, Amin Ul., Li, Jian Ping., Memon, Muhammad Hammad., Nazir, Shah., Sun, Ruinan. (2018). A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. Hindawi, GJCST, Volume:2018, pp: 01-22.
[19] Choubey, Dilip Kumar., Paul, Sanchita. (2016). GA_MLP NN: A Hybrid Intelligent System for Diabetes Disease Diagnosis. International Journal of Intelligent Systems and Applications (IJISA), MECS, Vol. 8, No. 1, pp. 49-59.
[20] Dangare, Chaitrali S., Apte, Sulabha S. (2012). Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. International Journal of Computer Applications, Volume 47– No.10, pp: 44-48, 2012.
[21] Frederix, Sankaran, S., Coninx, K., Dendale, P. (2016). MobileHeart, a mobile smartphone-based application that supports and monitors coronary artery disease patients during rehabilitation. IEEE, pp: 513-216.
[22] Bala, Kanchan., Choubey, Dilip Kumar., Paul, Sanchita. (2017). Soft Computing and Data Mining Techniques for Thunderstorms and Lightning Prediction: A Survey. International Conference of Electronics, Communication and Aerospace Technology (ICECA 2017), IEEE, during 20-22 April, 2017, ISBN 978-1-5090 5686-6, RVS Technical Campus, Coimbatore, Tamilnadu, India, Vol. 1, pp. 42-46.
[23] Anooj, P.K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University – Computer and Information Sciences, pp: 27-40.
[24] Kahramanli, Humar., Allahverdi, Novruz. (2007). Design of a hybrid system for the diabetes and heart diseases. Elsevier, pp: 82-89.
[25] Giardina, Marisol., Azuaje, Francisco., McCullagh, Paul., Harper, Roy. (2006). A Supervised Learning Approach to Predicting Coronary Heart Disease Complications in Type 2 Diabetes Mellitus Patients. IEEE.
[26] Xu, Shan., Zhang, Zhen., Wang, Daoxian., Hu, Junfeng., Duan, Xiaohui. (2017). Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework. IEEE, pp: 228-232.
[27] Bala, Kanchan., Choubey, Dilip Kumar., Paul, Sanchita., Lala, Mili Ghosh Nee. (2018). Classification Techniques for Thunderstorms and Lightning Prediction- A Survey. Soft Computing-Based Nonlinear Control Systems Design, IGI Global, ISBN 978-1-5225-3531-7, pp. 1-17.