International Journal of IT & Knowledge Management (ISSN: 0973-4414)
Volume 15 • Issue 2 pp. 57-62 Jan-June 2022 www.csjournals.com

A Relative Study of Heart Disease Prediction Using Machine Learning Techniques

Jaswinder Singh[1], Neha Bhadu[2]
Department of Computer Science and Engineering
Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India
jaswinder_singh_2k@rediffmail.com, nehabhadu100@gmail.com

ABSTRACT
One of the topmost reasons of death nowadays is heart disease. Healthcare professionals have always had a difficult time accurately predicting and detecting it. As a result, there is a necessity for a trustworthy, accurate, and practical method that can detect heart illness early on and provide the patient with an effective therapy before it progresses to a serious problem and, ultimately, to a heart attack. It has been shown that using machine learning (ML), it is possible to anticipate outcomes and make decisions based on the massive amounts of data produced by the healthcare industry.For heart disease prediction, diagnosis, and treatment throughout the past several years, a variety of available heart disease datasets have been subjected to machine learning (ML) algorithms and methodologies. This study investigates varied ML methodologies for the prediction of heart disease.
Keywords Machine learning (ML) · Cardiovascular disease (CVD) · Algorithms · Heart disease.

1. INTRODUCTION

The most essential part of the human circulatory system, which circulates blood containing oxygen to other body parts through arteries and veins, is the heart. Heart disease is the term used to describe any condition that affects the heart [1]. Cardiovascular diseases are another name for heart conditions (CVDs). There are numerous types of heart disease, including “deep vein thrombosis, pulmonary embolism, rheumatic heart disease, congenital heart disease, peripheral artery disease, cerebrovascular illness, coronary heart disease” and more.The “World Health Organization (WHO)” report claims that globally the main factor in death is heart disease. In 2019 around 17.9 million individuals worldwide had died of heart disease, which accounted for 32% of entire mortality, according to estimates. 85% of these deaths were from heart attacks and strokes [2].

The ability to take preventative action to lower this toll depends critically on the early, accurate, and effective medical detection of heart disease. Addressing behavioural risk factors can help prevent the majority of cardiac illnesses. For stroke and heart disease, the most crucial behavioural risk elements are poor eating, being inactive, smoking, and abusing alcohol. Being obese, having high blood sugar level, high BP and higher blood lipids are just a few symptoms that people may encounter as a result of behavioural risk factors. [3].Using invasive techniques to diagnose diseases is expensive, time-consuming, uncomfortable and because they are carried out by humans, they may produce false results. Therefore, a method that can non-invasively diagnose cardiac disease in less time and at a lower cost is needed in order to get better and faster results. One such method is ML.

ML is one of the fast expanding branches of AI. Its main objective is to construct systems, let them learn, and afterwards use what they've learnt to generate predictions. By lowering the errors in projected and actual outcomes, it is an alternative to conventional prediction modelling approach employing a computer to analyze complicated and non-linear interactions among many elements [4]. Huge amounts of data from many other sectors, including the vital field of medicine, can be analyzed using ML algorithms. In health sector, there is tremendous amount of patient information.The provision of patient-friendly, high-quality clinical services is a significant problem for healthcare companies. Good service delivery necessitates accurate patient diagnosis, identification of an appropriate treatment, and avoidance of incorrect diagnoses. Medical diagnosis should be efficient, trustworthy, and backed by computational methods to cut down the existent toll of identification tests. For this, several ML algorithms must mine the medical data.

Using various ML algorithms, many researchers had suggested various models to predict cardiac disease. The accuracy of disease prediction and outcome optimization are the primary issues facing researchers nowadays. The chief goal of this work is to carry out a comparative examination of various ML strategies put out by various researchers.

Following is the breakdown of the remaining portions of this paper: An overview of the several models for heart disease put forth by various researchers is found in Section 2. In Section 3, the various ML techniques analyzed in the literature are contrasted. Section 4 contains the conclusion.

2. RELATED WORK

Different ML methods have been utilized by researchers in order to create models for the prognosis of cardiac disease. This section discusses a few of them.

In [5] Xu S et al. (2017), provided a paradigm for enhanced random forest and CFS subset evaluation for heart disease risk prediction. "CFS subset evaluation" mechanism and the "best-first-search" method were compounded for feature selection in order to reduce dimension. The results of several tests and experiments of various kinds showed that the most accurate classifier available is random forest. The correctness and applicability of two separate datasets were verified.The system had a much greater accuracy of 91.6% than all the other techniques in the CHDD test. It fared better than some other ML classifiers in the "People's Hospital dataset" test, with the exclusion of SVM, with an accuracy of 97%. RF, however, only took half as long as SVM.

In [6] S. M. M. Hasan *et al.* (2018), compared the effectiveness of various classification methods including "KNN, Decision Tree (ID3), Gaussian Naive Bayes, Logistic Regression, and Random Forest" on dataset of heart disease. The classification accuracy for Logistic Regression was 92.76%, making it the best performing method.

In [7] Amin Ul Haq *et al.* (2018), developed an intelligent hybridized predictive model based on ML for the detection of cardiac disease. The system was evaluated using heart disease data from cleveland. "Relief, mRMR, and LASSO" algorithms for feature extraction were combined with seven well-known classifiers, including "LR, K-NN, ANN, SVM, NB, DT, and RF", to choose the critical features. When chosen using the FS algorithm Relief, it was discovered that LR with "10-fold cross-validation" demonstrated the maximum accurateness at 89%. SVM (linear) with mRMR has the highest level of specificity when compared to LR with LASSO's and Relief FS algorithms.On features chosen by Relief, the ANN (MLP) classifier with sixteen hidden neurons had the maximum sensitivity at 100%. Thus, it was discovered that FS algorithms might speed up processing while also increasing the classification accuracy of classifiers. Classifiers that use the Relief FS algorithm's crucial feature selection perform astonishingly well when measured against LASSO and mRMR.

In [8] S. Mohan *et al.* (2019), projected the "hybrid random forest with linear model as a technique (HRFLM)". The strategy sought to identify relevant features by utilizing ML algoritms, enhancing the accurateness in prediction of cardiovascular diseases. The initial version of the prediction model included a number of feature combinations and well-known classification techniques. The HRFLM's accuracy level was found to be 88.7%, resulting in an improved performance level.

In [9] A. N. Repaka *et al.* (2019), constructed "SHDP (Smart Heart Disease Prediction)" utilizing navies bayesian algorithm which forecasted risks for developing heart disease. The proposed technique includes several steps, including accumulation of dataset, user enrollment and logging in (according to the usage), categorization utilizing navies bayesian, prediction and assured data transmission utilizing "AES(Advanced Encryption Standard)". According to the results, the suggested method generated an accuracy of 89.77%.

In [10] S. Bashir *et al.* (2019), employed methods for selecting features to enhance the predictability of cardiac disease. For selecting the features "Minimum Redundancy Maximum Relevance Feature Selection (MRMR)" technique was used. The UCI heart disease dataset was then subjected to individual applications of the varied ML algorithms like DT, LR, naive bayes, Logistic Regression (SVM), and RF in Rapid Miner. MRMR/FS using logistic regression (SVM) was discovered to have the highest accuracy.

In [11] Shah, D. *et al.* (2020), used data from the UCI repository to predict cardiac disease utilizing supervised ML methodologies like DT, naive bayes, KNN, and RF. Only 14 of the 76 attributes were taken into account when conducting the experiment. Using the WEKA tool, the data were pre-processed. After developing four methods, it was discovered that KNN had the best accuracy.

In [12] J. P. Li. *et al.* (2020), introduced the "fast conditional mutual information (FCMIM)" feature selection algorithm to overcome the challenge of selecting features. The "Leave-One-Subject-Out Cross-Validation (LOSO)" method was used to select ideal hyper-parameters for selecting the best model. On a dataset of Cleveland heart disease, the suggested technique was tried. The suggested "FCMIM" is practicable with a SVM classifier for constructing a advanced intelligence system that can recognise cardiac illness, delivering an accuracy of 92.37%, after its performance was contrasted with that of cutting-edge existing methods.

In [13] A. Singh and R. Kumar (2020), calculated the predictive accurateness of four ML techniques, namely "KNN, DT, LR, and SVM". The algorithms were trained and tested using the UCI repository data for heart disease. Python programming was used to carry out the implementation with a Jupyter notebook. The findings demonstrated that KNN had the highest accuracy among all, coming in at 87%.

In [14] N.L. Fitriyani *et al.* (2020), integrated "DBSCAN, SMOTE-ENN, and XGBoost-based MLA" to construct an accurate "heart disease prediction model (HDPM)" for a ":clinical decision support system (CDSS)".The training data distribution was balanced using a mix of the "Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE-ENN)" method, "Density-Based Spatial Clustering of Applications with Noise (DBSCAN)" was utilized to find and remove outliers, and XGBoost to anticipate cardiac disease.The framework was constructed utilizing 2 publicly accessible datasets namely cleveland and staglog, and its output was compared to that of some other frameworks (naive bayes, LR, MLP, SVM, DT and RF) as well as the findings of earlier research. The accuracy rates for the datasets from statlog and cleveland for the HDPM, which outperformed other models, were 95.90% and 98.40%, respectively.

In [15] Katarya, R. and Meena, S.K. (2021), conducted a comparison of various ML techniques for the anticipation of cardiac disease on the UCI dataset. In terms of accuracy and other evaluation parameters, it was discovered that RF generated the optimum results.

In [16] Rani, P. *et al.* (2021), developed a hybridized system for decision support that aids in the early identification of cardiac disease and is based on the clinical characteristics of the patient.SVM, naive bayes, LR, RF, and adaboost classifiers were used to develop the model. With RF classifier, the system has proven to produce the most accurate results. The model was tested using the clevel and heart disease dataset from the UCI library. The outcomes showed that RF performed best, with an accuracy of 86.60%.

In [17] M. Kavitha *et al.* (2021), suggested a cutting-edge machine learning strategy to forecast cardiac problems. The heart disease dataset from cleveland was subjected to the use of RF, DT, and hybridized model (hybrid of RF and DT) machine learning approaches. The user's input parameter was required by the interface's design in order to forecast cardiac disease. According to the findings, the hybrid model achieved an accuracy rate of 88.7%.

In [18] Ali, Md. Mamun et al. (2021), utilized many supervised ML algorithms, and their efficacy for predicting cardiac disease was compared. The Kaggle website hosted the dataset. The study discovered that RF, with a 100% accuracy rate, outperformed KNN, DT, and other classification algorithms.

3. Major Findings

Table 1: Methodology used by different researchers in the prediction of heart disease

Author' Name, Year [Ref]	Dataset	Techniques Used	Tools	Accuracy
Xu, S. *et al.* (2017) [5]	1. Cleveland 2. People's Hospital	CFS Subset Evaluation with Improved RF	N/A	1. CHDD Test:91.6% 2. People's Hospital

	Dataset (PKU)	Framework and BFS		Dataset Test: 97%
S. M. M. Hasan *et al.* (2018) [6]	Cleveland	Gaussian Naive Bayes, DT (ID3), LR and RF	Anaconda Python (Spyder 3.6)	92.76% in Logistic regression
Amin, Ul Haq. *et al.* (2018) [7]	Cleveland	Feature Selection Using ANN, SVM, LR, KNN, NB, DT, and RF Relief, mRMR, and LASSO	N/A	Logistic regression with Relief: 89%
S. Mohan *et al.* (2019) [8]	Cleveland	"Linear Method (LM)", RF Combining a "Linear Model and Hybrid Random Forest (HRFLM)"	R Studio Rattle	HRFLM: 88.7%
A. N. Repaka *et al.* (2019) [9]	UCI Repository	Bayes Net, "MLP (Multi-Layer Perception)", Navies Bayesian, "SMO (Sequential Minimal Optimization)", "AES (Advanced Encryption Standard)", and "PHEA (Parallel Homomorphic Encryption Algorithm)"	N/A	Navies Bayesian: 89.77%
S. Bashir *et al.* (2019) [10]	UCI Repository	DT, LR, Logistic regression SVM, Naïve Bayes and RF "Minimum Redundancy Maximum Relevance Feature Selection (MRMR)"	Rapid Miner	MRMR/FS with Logistic Regression (SVM): 84.85%
Shah, D. *et al.* (2020) [11]	Cleveland	RF, KNN, DT, and Naive Bayes	WEKA and Python	KNN: 90.78%
J. P. Li. *et al.* (2020) [12]	Cleveland	mRMR, LASSO, NB Relief, KNN, SVM, ANN, LR, DT and "Local-learning-basedfeatures-selection (LLBFS)"	N/A	FCMIM-SVM: 92.37%

		"Leave-one-subject-out cross-validation (LOSO)" "Fast conditional mutual information (FCMIM)" feature selection algorithm		
A. Singh and R. Kumar (2020) [13]	UCI Repository	KNN, LR, DT and SVM	Python Jupytor Notebook	KNN: 87%
N. L. Fitriyani et al. (2020) [14]	Statlog and Cleveland	NB, LR, MLP, SVM, DT, RF and "heart disease prediction model (HDPM)"	Python V3.6.5	HDPM for Statlog: 95.90% HDPM for Cleveland: 98.40%
Katarya, R. and Meena, S.K. (2021) [15]	UCI Repository	LR, Naive Bayes,KNN,SVM, DT, RF, ANN, DNN and MLP	N/A	Random Forest: 95.60%
Rani, P. *et al.* (2021) [16]	Cleveland	SVM, NB, LR, random forest, and adaboost classifiers "Hybrid Genetic Algorithm (GA) " and "recursive feature elimination " for hybridized feature selection "SMOTE (Synthetic Minority Oversampling Technique)"	Python	Hybrid Decision Support System: 86.60%
M. Kavitha *et al.* (2021) [17]	Cleveland	RF, DT and hybrid of RF and DT	Python 3.7.6	Hybrid Model: 88.70%
Ali, Md. Mamun *et al.* (2021) [18]	Kaggle	KNN, DT and RF	Weka3.8.3 Python 3.8.5	Random Forest: 100%

CONCLUSION

Medical professionals have discussed the use of ML, particularly for predicting cardiac disease. To increase algorithm precision and produce findings that can be trusted, more precise feature selection techniques are applied. If a precise form of cardiac disease is identified, the patient should receive care for that particular illness. In essence, it can be said that a dataset of adequate samples and trustworthy data will be used to develop a heart disease prediction model. The dataset must be pre-processed as a result, as pre-processing is the most essential step in getting the dataset ready for the ML algorithm and improving results. To form a prediction

model that produces precise results, the accurate algorithm be required to be engaged. Finally, using machine learning to identify cardiac disease is a crucial task that can help patients as well as healthcare professionals.

References

1. Rani, P., Kumar, R., Ahmed, N.M.O.S. et al. A decision support system for heart disease prediction based upon machine learning. J Reliable Intell Environ 7, 263–275 (2021). https://doi.org/10.1007/s40860-021-00133-6.
2. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
3. https://www.who.int/health-topics/cardiovascular-diseases.
4. Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020). https://doi.org/10.1007/s42979-020-00365-y.
5. Xu S, Zhang Z, Wang D, Hu J, Duan X, Zhu T. "Cardiovascular risk prediction method based on CFS subset evaluation and random forest classifcation framework," 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing. 2017. p. 228–232. https://doi.org/10.1109/ICBDA.2017.8078813.
6. S. M. M. Hasan, M. A. Mamun, M. P. Uddin and M. A. Hossain, "Comparative Analysis of Classification Approaches for Heart Disease Prediction," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018, pp. 1-4, doi: 10.1109/IC4ME2.2018.8465594.
7. Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Mobile Information Systems, vol. 2018, Article ID 3860146, 21 pages, 2018. https://doi.org/10.1155/2018/3860146.
8. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
9. A. N. Repaka, S. D. Ravikanti and R. G. Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 292-297, doi: 10.1109/ICOEI.2019.8862604.
10. S. Bashir, Z. S. Khan, F. Hassan Khan, A. Anjum and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches," 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 2019, pp. 619-623, doi: 10.1109/IBCAST.2019.8667106.
11. Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020). https://doi.org/10.1007/s42979-020-00365-y.
12. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.
13. A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.
14. N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511.
15. Katarya, R., Meena, S.K. Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. Health Technol. 11, 87–97 (2021). https://doi.org/10.1007/s12553-020-00505-7.
16. Rani, P., Kumar, R., Ahmed, N.M.O.S. et al. A decision support system for heart disease prediction based upon machine learning. J Reliable Intell Environ 7, 263–275 (2021). https://doi.org/10.1007/s40860-021-00133-6.
17. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.
18. Ali, Md. Mamun & Paul, Bikash Kumar & Ahmed, Kawsar & Bui, Francis & Quinn, Julian & Moni, Mohammad Ali. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology and Medicine. 136. 104672. 10.1016/j.compbiomed.2021.104672.