

Enhancing Cardiovascular Disease Prediction with Machine Learning: A Comparative Study Using the UCI Heart Disease Dataset

Hanwen Li^{ID}^a

College of Information Science, University of Arizona, Arizona, U.S.A.

Keywords: Cardiovascular Disease, Machine Learning, UCI Heart Disease, Model Evaluation.

Abstract: At present, cardiovascular diseases (CVDs) remain the principal cause of death at large global scale, so early detection is essential for improving patient outcomes. In this study, machine learning (ML) techniques are introduced in an effort to oppress heart disease prediction work, using the University of California, Irvine (UCI) Heart Disease Archive Database as the facility for model evaluation. Thesis want to talk about the performance of a number of ML models: Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, Artificial Neural Networks (ANN) and Deep Neural Networks (DNN). The research involves data preprocessing steps including normalization and imputation steps of the dataset, followed by training and testing models to evaluate accuracy and precision. The data source consists of 303 instances with 14 angles. Logistic Regression achieves the highest accuracy at 93.40 percent. ANNs and DNNs show strong capabilities for pattern recognition but also encounter overfitting problems. Decision Trees provide valuable interpretation but have only moderate generalization power. Performance is precisely sensitive to data characteristics. These findings illustrate the potential of ML techniques as a means to amplify heart disease prediction (providing clinicians with even more accurate tools for early diagnosis and personalized treatments) not only their strength but also conclusion on real-time where direct benefits will certainly be received by patients ourselves within health care settings.

SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Globally, heart disease is the number one killer annually claiming 17.9 million lives, according to World Health Organization figures from the past year. The silent but intricate progression of heart problems means that it is important to have advanced early warning instruments such as electrocardiographs on hand. Thus, it's not enough to simply prolong patients' lives; professionals need to manage the immense burden of care and continuously improve healthcare systems. Innovative technologies in cardiovascular diagnostics have included the integration of machine learning (ML). This integration makes use of the robust data analysis capabilities of ML algorithms to increase precision in detection, and helps create custom therapeutic strategies (Cardiovascular, 2021; Singh and Kumar, 2020; Bhavkar et al., 2024; Katarya and Meena, 2020).

The use of machine learning (ML) has shown its irreplaceable role in the prediction of heart diseases, and it was demonstrated to be effective in different clinical studies. A few studies have examined ML algorithms such as decision trees, K-nearest neighbors (KNN), support vector machine (SVM) and neural network by Singh & Kumar et al. (Enad and Mohammed, 2023; Louridi et al., 2021). These research works signed off on approaches which combine several algorithms, such as Random Forest or hybrid models that result in better prediction accuracy and higher efficiency as compared to others. The research underscores the importance of multiple algorithmic strategies for handling patient-specific information when calibrating diagnostic performance. Moreover, novel techniques such as deep learning and active learning are also expanding the horizons for cardiovascular diagnostics. It does not show the first wave of Corona Virus Disease (COVID)-19 cases in India, but given the enormous population living there, even a sliver of early insight

^a <https://orcid.org/0009-0003-2700-7595>

can help to save lives. These techniques are highly effective for intricate patterns within large datasets (Bhavekar et al., 2024; Louridi et al., 2021), with electrocardiogram (ECG) analytics further enhancing chances of early detection and patient survival. These essentially innovative strategies focusing on the exploitation of the most informative data segments, significantly diminished the actual number of samples that needed to be trained while maintaining high accuracy in regard to predictive models in cardiology (Srinivasan et al., 2023).

Current efforts to exploit cutting-edge Artificial Intelligence (AI) methods, such as quantum machine learning and AI-enabled analytical tools, have increased the predictive capability of heart disease (Enad and Mohammed, 2023; Martinez et al., 2022; Schepart et al., 2023). Research by Kumar et al. (Enad and Mohammed, 2023) applied quantum machine learning being more efficient than classical computing to deal with large computational demands, achieving diagnoses faster and more accurate than conventional approaches. In addition, the use of artificial intelligence in electronic health record (EHR) data analysis and medical imaging is changing personalized treatment and disease progression prediction such as in the area of precision medicine (Martinez et al., 2022; Schepart et al., 2023; Absar et al., 2022; Lambay et al., 2024).

The main aim of this study is to do comparative analysis and evaluation of the performance among machine learning techniques regarding the prediction of cardiovascular diseases (CVDs). Here, this study explores the use of key ML technologies in cardiovascular prediction, presenting a comprehensive review of major models, methods and implementation paradigms in this area. These advances and challenges in cardiovascular diagnostics have applications to both pulmonary aspects (highlighting the potential of ML in enhancing early detection and patient-specific approaches to care). The insights ascertain further understanding of current applications for ML and hypostatize the doorway to potential future and ongoing innovations in predictive healthcare. Discussion is subsequently provided related to the strengths and weaknesses of these methods, and the directions for technological improvements.

2 METHODOLOGY

2.1 Dataset Description

University of California, Irvine (UCI) Heart Disease Dataset: The ground truth dataset for classification of cardiovascular disease, consists of records from 303 patients with 14 attributes (age, sex, etc.), which are important for prediction the heart conditions (Janosi et al., 1988). This dataset, sourced from well-known authorities like Cleveland Clinic and Hungarian Institute of Cardiology, offers a very rich relational database on which virtually every type of machine learning model can be applied. These include Logistic Regression, Random Forests, and Support Vector Machines that help in discovering complex cardiovascular patterns. Additionally, several ML algorithms have been developed to improve the pattern recognition ability in heart disease data using this dataset. This holistic model of care emphasizes the importance of precision medicine and risk stratification in cardiology, essential for individualized therapeutic approach tailoring and raising patient management to higher standards. The following sections will unpack these methodologies further, shedding light on possible developments in using ML to improve cardiac health.

2.2 Proposed Approach

This study aimed to update the prediction of cardiovascular diseases (CVDs) by means of ML technologies, starting with an outline on how ML more efficiently handles complicated data sets as compared to conventional approaches in cardiology. Algorithms are used to locate certain patterns and risk factors in patient data that may indicate some form of heart disease. It starts with an in-depth investigation of existing ML solutions and their roles for heart disease diagnostics, where a review is made on various diagnostic algorithms such as Logistic Regression, Random Forests and Support Vector Machines along with strengths/weaknesses related to processing cardiovascular data. The first part of this research is data pre-processing. This introduced UCI Heart Disease Dataset to fill in the missing values and scaling the numerical attribute set of attributes (Janosi et al., 1988).

After preprocessing the data, research examines the performance of common ML models over this dataset and checks its predictive power. By comparing models, the study leads to finding what model scores best for heart disease prediction. The study then proceeds to more advanced methodologies

such as deep learning for better model performance while extracting intricate information from ECGs and imaging data. The objective for cardiovascular healthcare is to refine ML models that enable improved early detection, and personalized treatments will lead to better patient outcomes and save lives at an even lower cost. The discussion and conclusion sections critically review the findings regarding the strengths and limitations of using ML in cardiovascular diagnostics. Moreover, those sections scrutinized the clinical impact of practical applications using created models as well as new pathways for further studies. Illustration of the research process is in Figure 1.

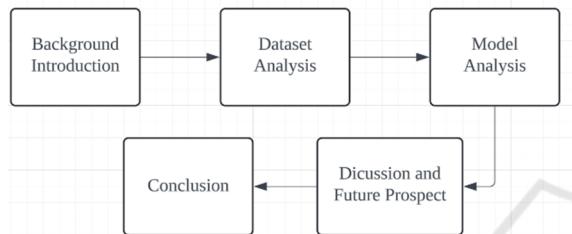


Figure 1: The flowchart of the survey (Picture credit: Original).

2.2.1 Introduction of ML

Machine learning teaches computers to learn from data to make informed decisions on their own, allowing systems to be trained and optimized without human intervention. This process is how humans also learn to view a data set and detect patterns based on multiple instances of this experience. In healthcare, machine learning helps to parse through patient data of variables (e.g. age, blood pressure, cholesterol) and predict individual health outcomes or potential future events like developing a disease. The present in-depth review aims at investigating a variety of ML models specialized for CVD prediction. These models are designed to perform well when working with large datasets and provide an accurate prediction of heart disease from a combination of risk factors.

2.2.2 Logistic Regression

Logistic Regression is a well-known supervised learning technique that is applied to binary classification problems like heart-disease prediction. This algorithm provides the probability estimates of heart disease risk using a sigmoid or logistic function, which ensures that this ranges from 0 to 1 by transforming a linear combination (weighted sum) of observed predictors like age, blood pressure and cholesterol levels. Of particular note is its simplicity

and transparency that enable a straightforward interpretation regarding how different risk factors relate to probability of disease occurrence. It has the advantage but adapts essentially by drawing a straight line through all input features and the logarithmic odds of the result; this can be too simple to capture more complex data patterns. Katarya and Meena (Katarya and Meena, 2020) reported its usefulness in predicting heart diseases due to the computational efficiency of this algorithm on binary outcomes. Augmentations such as regularization are proposed to improve its precision by preventing overfitting when dealing with complex high-dimensional datasets (Katarya and Meena, 2020).

2.2.3 Decision Trees

Heart disease is usually predicted using decision trees because of their simple structure and interpretability. By considering input features, they split the dataset into finer and finer pieces to form a tree-like structure in these models. In this context, every internal node represents a certain feature (for example, whether the grain is green) and the branches are the outcomes of that decision (yes/no), with leaf nodes showing outcome. Decision Trees are transparent enough to be directly utilized in clinical settings, so the extracted rules can easily be executed by clinicians. However, as highlighted by Singh and Kumar (Singh and Kumar, 2020), these models are prone to overfitting particularly in the presence of small or noisy datasets. This situation is known as overfitting and happens when the model too much tailor itself to the peculiarities of the training data, which makes it become less effective in others new datasets not yet seen.

2.2.4 K-Nearest Neighbor

K-Nearest Neighbors is a flexible algorithm used for classification and regression problems, including predicting heart disease (Singh and Kumar, 2020). The algorithm works in a way that it classifies any new data points based on the majority class of 'k' closest data point within the dataset. For example, in medical diagnosis of heart disease (the target), the algorithm measures how close a combination of attributes is to all other data points (patients) and labels the data point with the most frequent diagnosis of its nearest neighbors. KNN has a number of advantages over other model which includes that it does not make any assumption about the data distribution, so it is non-parametric and can be used across datasets. However, KNN can be computationally intensive especially for larger

datasets which entail calculating distances from each datapoint to all data points as pointed out by Singh and Kumar (Singh and Kumar, 2020). Furthermore, KNN has a high sensitivity to the value of the parameter 'k', and it can be easily reduced by noise in data with low prediction accuracy.

2.2.5 ANN & DNN

Artificial Neural Networks (ANNs) and their advanced variation, Deep Neural Networks (DNNs), have become important tools in understanding the prediction of heart disease because of their nature of capturing complex non-linear relationships within data (Martinez et al, 2022; Schepart et al., 2023). ANNs work by layers of nodes that are interconnected which imitates how human brain operates and this allows ANN to process certain observations like age, cholesterol or blood pressure. These networks are especially adept at identifying subtle relationships between risk factors in a heart disease model that traditional algorithms might overlook due to the sheer number of dimensions. Deep learning networks (DNNs) are essentially ANNs with more than one hidden layer that allow the network to learn highly abstract data concepts. This trait in DNNs has made them more apt for dealing with huge and complex data as they are more capable of extracting richer meaning out of the jumbled information. Existing works have extensively proven that DNNs yield great performance when it comes to heart disease prediction in large datasets (Schepart et al., 2023). However, these models tend to be computationally expensive and prone to overfitting, particularly when exposed to noise, small sample size or complex model (Ying, 2019). In order to manage this problem of overfitting, techniques like dropout-- disables some random neurons during training or regularization--are used. Despite this limitation, ANNs and DNNs are indispensable as heart disease prediction models because they can generalize from complicated data structures and help to create personalized medical intervention.

3 RESULT AND DISCUSSION

3.1 The performance of models

The effectiveness of different machine learning algorithms to predict cardiovascular diseases is reported in this study as shown in Table 1 above. The data for this analysis comes with UCI Heart Disease dataset which contains 303 instances and 14

important attributes (age, cholesterol levels, maximum heart rate etc.), all of which are very significant indicators of the risk from heart disease. The primary evaluation metrics for machine learning models were accuracy, representing the correctness of the predictions overall, and precision, illustrating how many true positive cases can be found in all predicted as positive.

Table 1: Comparison of various techniques (Katarya and Meena, 2020).

Techniques	Accuracy (%)	Precision
Logistic Regression	93.40	0.4589
KNN	71.42	0.4770
Decision Tree	81.31	0.5
ANN	92.30	0.4524
DNN	76.92	0.5

Logistic Regression emerged as the top performer with the highest accuracy 93.40%. Nevertheless, its accuracy was quite low at 0.4589, meaning that the model could possibly generate false positives more. In contrast to this, KNN having the lowest accuracy of 71.42% as it is sensitive to 'k' selection and handling nature of dataset which has been known in noisy or complex data. As token tabled 1 it achieved an accuracy of 81.31%, the precision score was 0.5 by Decision Tree model. The relatively simple and interpretable nature of this model would seem to make it a practical choice for use within clinical practice. Its well-calibrated accuracy demonstrates only mild overfitting and seemingly modest ability to accurately pick up true cases of heart disease.

Similarly, ANNs also performed with as high accuracy as Logistic Regression, 92.30%. However, its precision was slightly less at 0.4524 (it probably has a chance of overfitting smaller datasets so ends up labelling more stuff false positives). While it was less accurate than Logistic Regression and ANNs, the DNN model is especially helpful for large datasets where complicated, non-linear data patterns can be captured. DNNs are so precise that they have a mild power of at most correctly predicting true positive cases.

3.2 Discussion

The insights gained by experiment in heart disease prediction are based on several machine Learning models along with corresponding model selection, clinical applicability and some considerations for future work. Every algorithm has its own advantages

and disadvantages. Logistic Regression is known for its simplicity and scalability, making it very appealing in the medical field where quick decisions need to be taken. However, with linear regression all the complex health data cannot be modelled. Improvements such as iterating it with other regularization methods that aim to increase accuracy and precision. In contrast, KNN lacks high-dimensional or unbalanced datasets, which inhibits its performance for complex clinical cases. Possibly, applying optimization techniques and data preprocessing steps, such as normalization, can improve its robustness and scalability for health.

As the outcome must be clear in clinical settings for decision making, Decision Trees are often favored due to their interpretability. It makes healthcare workers comfortable as they can pick up the model easily. The major issue with Decision Trees is that they are prone to overfit, particularly on smaller or noise datasets, and not as good at generalization. Despite a relatively low accuracy and precision, their good interpretability is still an advantage of such models when used in clinical practice. One of the biggest hurdles in predicting heart disease has been finding a trade-off between model complexity and interpretability necessary for clinical decisions. However, DNNs, while often being able to achieve very high accuracies, are still considered "black boxes" in that they leave decisions unexplained and as such are not well-suited for medical applications where interpretability is of utmost importance. Logistic Regression and Decision Trees are simpler models that give better transparency, but they may not catch the more complicated patterns in the data. Further research employing hybrid models that amalgamate the best of both types and utilizing explainable AI (XAI) methods to, in general, increase transparency of more complex models while keeping the balance with accuracy can be pursued.

Moreover, data imbalance is common in medical datasets because typically the number of the healthy is greater than the number of the sick and this also leads to deteriorating on model performance. Over-sampling techniques like Smote (Synthetic Minority Over-sampling Technique) or cost-sensitive learning can help balance these classes which in turn will increase the sensitivity and precision of heart disease detection. Lastly, future work for ML models should seek the integration of them into clinical decision support systems. Integrating predictive models into front-line clinical systems would streamline diagnostic workflows and help clinicians pinpoint high-risk patients, readmissions, or alarms for time-sensitive decisions. The extent of collaboration

between data scientists and medical professionals will also be vital in order to ensure that models are not only accurate but useful for actual clinical applications.

4 CONCLUSIONS

This study explores how different machine learning models would perform to predict cardiovascular disease using the UCI Heart Disease dataset. This research was performed by using models such as Logistic Regression, KNN, Decision Trees, ANN and DNN to study the significant risk factors Responsible for predicting heart diseases. The steps include extensive data preprocessing, followed by an evaluation from metrics such as accuracy and precision. Results obtained from the experiments showed that Logistic Regression was the most accurate, but ANN/DNN had better pattern recognition ability in spite of suffering overfitting problems. Decision Trees provided good interpretability yet had a problem in generalization, and performance was dependent on data characteristics. Further work should strive to improve the generalizability of these models and to tackle the overfitting and data imbalance problems. Algorithms and strategies like data augmentation or transfer learning will be investigated in order to get more out of neural networks. Secondly, thesis will explore the integration of XAI methods to further improve model interpretability and render these state-of-the-art techniques more suitable for clinical use. It will also be necessary to work on increasing the % accuracy and reduce false positives even more before these methods are ready for use in practical health facilities.

REFERENCES

- Absar, N., Das, E. K., Shoma, S. N., Khandaker, M. U., Miraz, M. H., Faruque, M. R. I., Tamam, N., Sulieman, A., & Pathan, R. K., 2022. The Efficacy of Machine-Learning-Supported Smart System for Heart Disease Prediction. *Healthcare*, 10(6), 1137.
- Bhavekar, G.S., Das Goswami, A., Vasantrao, C.P., et al. 2024. Heart disease prediction using machine learning, deep Learning and optimization techniques-A semantic review. *Multimedia Tools and Applications*, 1-28.
- Cardiovascular. D., 2021. Retrieved on 2024, Retrieved from:[https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
- Enad, H., & Mohammed, M., 2023. A Review on Artificial Intelligence and Quantum Machine Learning for Heart Disease Diagnosis: Current Techniques, Challenges

- and Issues, Recent Developments, and Future Directions. 11. 08-25.
- Janosi, A., Steinbrunn, W., Pfisterer, M., et al. 1988. Heart disease. uci machine learning repository. UCI Machine Learning Repository.
- Katarya, R., & Meena, S.K., 2020. Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. *Health and Technology*, 11, 87 - 97.
- Lambay, M.A., Mohideen, S.P., 2024. Applying data science approach to predicting diseases and recommending drugs in healthcare using machine learning models – A cardio disease case study. *Multimed Tools Appl*, 83, 68341–68361.
- Louridi, N., Douzi, S., & El Ouahidi, B., 2021. Machine learning-based identification of patients with a cardiovascular defect. *J Big Data*, 8, 133.
- Martinez, D. S., Noseworthy, P. A., Akbilgic, O., Herrmann, J., Ruddy, K. J., Hamid, A., Maddula, R., Singh, A., Davis, R., Gunturkun, F., Jefferies, J. L., & Brown, S. A., 2022. Artificial intelligence opportunities in cardio-oncology: Overview with spotlight on electrocardiography. *American heart journal plus: cardiology research and practice*, 15, 100129.
- Schepart, A., Burton, A., Durkin, L., Fuller, A., Charap, E., Bhambri, R., & Ahmad, F. S., 2023. Artificial intelligence-enabled tools in cardiovascular medicine: A survey of current use, perceptions, and challenges. *Cardiovascular digital health journal*, 4(3), 101–110.
- Singh, A., Kumar, R., 2020. Heart disease prediction using machine learning algorithms. International conference on electrical and electronics engineering, 452-457.
- Srinivasan, S., Gunasekaran, S., Mathivanan, S.K. et al. 2023. An active learning machine technique-based prediction of cardiovascular heart disease from UCI-repository database. *Sci Rep*, 13, 13588.
- Ying, X., 2019. An overview of overfitting and its solutions. *Journal of physics: Conference series*. IOP Publishing, 1168: 022022.