

AI-Based Methods of Cardiovascular Disease Prediction and Analysis

Yifei Wang ^a

School of Physics and Astronomy, University of Edinburgh, Edinburgh, U.K.

Keywords: Cardiovascular Diseases, Machine Learning, Random Forest, Naïve Bayes, Learning Vector Quantization.

Abstract: Cardiovascular diseases (CVDs) remain a leading cause of global mortality, fuelling extensive medical research databases. Various models have been developed to predict CVDs from existing data, with machine learning (ML) emerging as a particularly effective method. This paper offers an overview of ML methods' performance in CVD prediction, specifically focusing on Random Forest (RF), Learning Vector Quantization (LVQ), and Naive Bayes (NB). Discrepancies among studies highlight the influence of factors such as data preprocessing, database selection, and sample size on ML performance. Consequently, determining the optimal ML method is challenging. This study lays the groundwork for future research, aiming to explore how each factor affects ML performance and facilitate improvements in subsequent studies. Furthermore, it encourages reproducibility through comprehensive literature review guidance. This paper lays foundation on future research into detailed influence of each factor on the performance of ML, and helps potential improvement for future studies. Reproductions are also hoped to be done with the guide of searched literatures.


1 INTRODUCTION

Cardiovascular diseases (CVDs) is a common term used to describe problems in heart and blood vessels. It is a leading reason of fatality across the world. Approximately 17.9 million people succumbed to CVDs in 2019, constituting 32% of the total global mortality, among which 85% were attributed to heart attacks and strokes (Gaziano, 2006). One challenging but important aspect of CVDs is to predict them. In fact, there is a huge amount of medical data contributed by CVDs, but it is not well-utilized to produce the desired result for patients (Gour, 2022). Therefore, it is well-motivated to come up with methods that predicts CVDs based on the pre-existing data sets.

Traditionally, models are manually built to predict the risk of having CVDs. Several well-known models include Qrisk, Framingham risk score, and score. Various complications and results of CVDs include death, heart stroke, heart failure and others, are together known as the major acute cerebrovascular and cardiovascular events (MACCE). MACCE is commonly used as the outcome of predictions performed by these models. However, among all the proposed models, considerable difference is observed

in the definitions of outcome. Most of the prediction models focus on CVDs, but some focus on event such as atrial fibrillation, stroke, e.t.c. The usage of potential causing factors, or predictors, in CVDs predictions also varies to a great extent among different researches. Most of the existing models includes age, smoking or not, blood cholesterol measurements, blood pressure, e.t.c. as predictors. Other prevalent predictors are diabetes and body mass index (BMI). Besides this heterogeneity, the number of models is already overwhelming, but most of them are not externally validated (Damen, 2016).

Recently, artificial intelligence (AI) has emerged as an effective tool for the prediction and diagnoses of CVDs. AI is a group of computational approached that mimic the way human learn, reason and solve problem. They can learn from the environment without being explicitly taught what to do, whereas the environment typically relates to a system that allows the interaction of the machine. In terms of CVDs predictions, the environment can be the CVD database. One branch of AI is machine learning (ML). ML can be mainly divided into three types: supervised, unsupervised and reinforcement learning. Supervised learning learns from data that is manually labelled and aims to gain the ability to make

^a <https://orcid.org/0009-0005-0600-3127>

predictions from the information given by the labels. On the contrary, unsupervised learning learns from fed information that is not labelled, trying to find the pattern behind it. Reinforcement learning (RL) machines interact and learn from the feedback, typically rewards, from the environment. The goal is to maximize the reward and therefore optimize the behaviour of machine. ML has been proved to be faster and more adaptive in handling meta-data compared with the traditional models (Damen, 2016). The primary aim of the paper is to offer an overview of various AI-based methods for predicting CVDs. Three algorithms are introduced and discussed: two traditional ML models, namely Random Forest (RF) and Naive Bayes (NB), along with Learning Vector Quantization (LVQ). The performance of ML models can be evaluated in numerous ways, with one common metric being the statistical accuracy of the model in predicting CVDs. Results from the implementations of these models are then presented for comparison.

2 METHODOLOGIES

2.1 Dataset Description

There are many datasets of CVDs, and the use of different datasets can potentially influence the outcome of training behaviour of a ML model. The most relevant dataset of this paper is the UCI machine learning repository (Anderies, 2022). There are four databases of UCI repository: Hungary, Cleveland, the VA Long Beach, and Switzerland. The database contains 76 attributes. Among these attributes, only a subset of 14 has been used in all the published researches. The dataset includes attributes like age, sex, chest pain type, resting blood pressure, e.t.c. Some nominal attributes have several types of classifications, such as the chest pain type. It includes four types of chest pains. The classifications of each attribute are represented by some numbers that can be found on the UCI-Cleveland database website. The other type of attributes is known as the numeric type, like the age of patients. Particularly, the Cleveland database is only one that has been used in purpose of ML.

There are two other possible sources of data, the Mendeley database and the Kaggle database. These databases have overlaps. The UCI-Cleveland database can also be found on Kaggle (Mendeley, 2024)(CHERNGS, 2020).

2.2 Proposed Approach

The approaches may vary across different implementations due to varying research designs or objectives. However, in general, the dataset is typically divided into two groups: the training dataset and the testing dataset. The machine learning algorithm is then applied to the training dataset to initiate the training process. Subsequently, the model's performance is evaluated using the testing dataset, utilizing various statistical metrics, and resulting in the final trained model. The overall pipeline of the proposed approach is illustrated in Figure 1. Commonly used statistical descriptions include precision and accuracy, defined by (true or false) positives and (true or false) negatives. Precision represents the rate of true positives among all positive results, while accuracy represents the overall rate of correct predictions. This paper emphasizes contribution to synthesizing and presenting a comprehensive overview of the methodologies employed in this domain, shedding light on the intricacies and variations in their implementation. Moreover, this paper highlights the significance of proposed approach in providing a structured framework for evaluating and comparing the performance of machine learning algorithms in predicting CVDs.

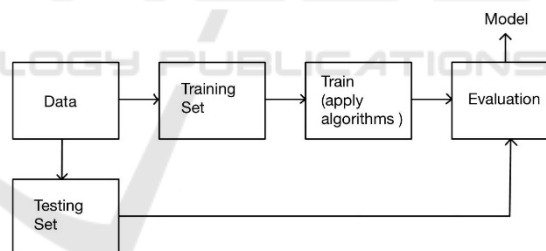


Figure 1: General pipeline of proposed models (Picture credit: Original).

2.2.1 Random Forest (RF)

Random forest is a useful ML method for many tasks, typically classification and regression. It is a combination of many tree predictors. These trees are random vectors from the same overall distribution, forming a forest. The selection process of each tree is independent of the others. As the number of trees in the forest increases, the generalization error of the forest converges almost surely to a limit. The correlation between the tree classifiers, together with the importance, or weight, of individual trees, decide the generalization errors of them. The training algorithm of RF applies a technique named bootstrap

aggregating, or bagging. In general, bagging repeatedly selects a sample randomly from the training set and fits trees to these samples. Then, each of the trees are trained on each bootstrapped subset of data. Because the subsets could vary between one another, each trained trees, or model, could also have variations. Finally, all the trained trees are aggregated together to reach an overall model. For regressions, the aggregation process is averaging, whereas for classifications, the overall model is combined through majority voting.

2.2.2 Naïve Bayes (NB)

Bayes' theorem is a fundamental theorem that describes the probability of an event. Mathematically, it is defined in the following way:

$$P(A), \text{ given } B = \frac{P(B), \text{ given } A}{P(B)} P(A) \quad (1)$$

where $P(B)$ is the probability of set B prior to evaluation, and $P(A)$ is the probability of set A prior to evaluation. The 'given' condition constrains the corresponding probability. $P(A), \text{ given } B$ is the posterior probability of A , and $P(B), \text{ given } A$ is the likelihood of B given that A happen.

NB is an algorithm suitable for classifications based on probability and Bayes' theorem. It is a type of supervised learning that assumes substantial independence of predictors. In other words, the given attributes should be independent of each other, allowing a simplification in calculation of probabilities given the Bayes' theorem. It is consequently naïve, because any dependency of the predictors could arbitrarily improve the probability. NB is widely used for text classification, spam filtering, and other applications where the input data consists of categorical or textual features.

2.2.3 Learning Vector Quantization (LVQ)

LVQ is a type of supervised learning algorithm based on competitions using artificial neural networks. It is suitable for classification tasks. LVQ is particularly useful for pattern recognition and classification tasks, especially when dealing with high-dimensional data. LVQ basically organizes the underlying pattern of descriptors into groups, and each of the groups consists of a transfer function. A group of learning patterns, together with recognized classifications a preparatory assignment of the output variable, will be passed to the system because LVQ uses a learning algorithm. After training, prototypes are used to measure the similarity between a given data point and

the different classes. The similarity is often computed using a distance metric, such as Euclidean distance, between the data point and each prototype vector (PV). Based on the evaluated similarity, an input vector is assigned to the class that is represented by the nearest PV. This assignment is used for classification purposes. The architecture of LVQ is shown in Figure 2.

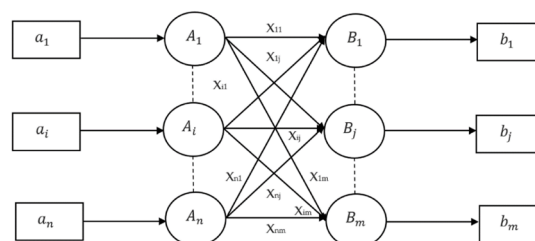


Figure 2: The architecture of LVQ algorithm is CVD-classification (Srinivasan, 2023).

There are n inputs from a_1 to a_n , and m outputs from b_1 to b_m . The neuron networks are fully attached to one another, and there are individual weights assigned to each of them.

2.3 Other Details

Most of the relevant researches perform multiple applications of the ML methods using different databases. It is possible that different database could cause potential difference in the learning behavior of ML methods. Therefore, only those using the UCI-Cleveland database are considered together as a comparison, and RF, NB and LVQ methods are mainly considered among all the other ML methods. Another important factor that causes potential difference on the training behavior is the pre-processing of data, such as the method of data-sampling from the database.

3 RESULTS AND DISCUSSION

In the experiment conducted by Saravanan Srinivasan et al., LVQ is proposed to be applicable in CVDs prediction as a new method (Srinivasan, 2023). Many ML methods are applied to the UCI Cleveland database, and the performance of LVQ is compared with various of different methods, including RF and NB. The result is illustrated in Figure 3.

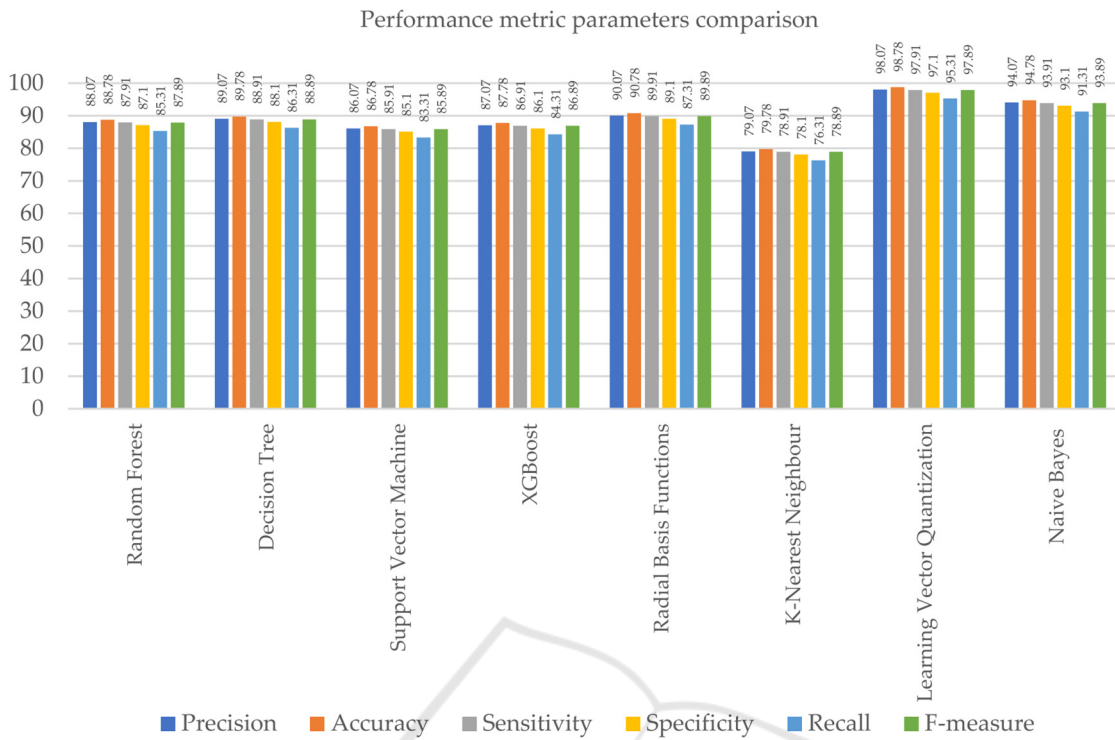


Figure 3: Graphical illustration of performance of various ML methods by Saravanan Srinivasan et al. (Srinivasan, 2023).

There are four other performance descriptions: sensitivity, specificity, recall and F-measure. Sensitivity is the true-positive rate given that the target is truly positive. On the other hand, specificity is the true-negative rate, given that the target is truly negative (Altman, 1994). Recall is also known as sensitivity. F-measure is a statistical measure of performance in prediction given by:

$$F - measure = \frac{2 * precision * recall}{precision + recall} \quad (2)$$

From the results, it is clear that the NB and LVQ method have considerably higher performance descriptions than the RF method. They also outperform the other methods that are not discussed.

In addition, LVQ has the highest categorization performance in forecasting CVDs with an accuracy of 98.07%, 4% higher than the NB method.

However, in the paper by Senthilkumar Mohan et al., the result is largely different. They use the same UCI-Cleveland database, and perform several ML for CVDs predictions. From Figure 4, one significant difference is that the performance parameter varies largely for each of the ML method, especially for the specificity that scores considerably smaller than the other parameters.

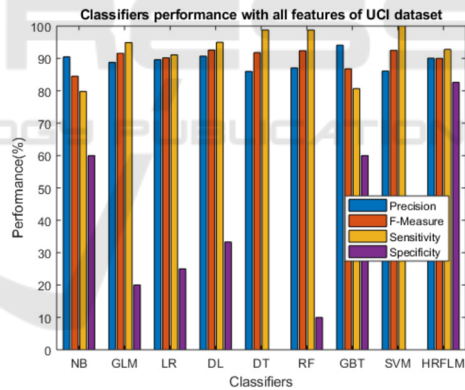


Figure 4: Graphical illustration of performance of various ML methods by Senthilkumar Mohan et al. (Kiran, 2022).

What’s more, the NB and RF method here has smaller discrepancies in precision, but the F-measure and sensitivity of RF is significantly larger that that of NB. The RF method is therefore concluded to be of better performance here. The HRFLM method (the rightmost one is Figure 4) is a method proposed by Mohan et al. based on RF. It combines RF with linear model that assumes a linearly-separable property of data and tries to learn the weight of separated data. The achieved performance is the best among all the other ML methods in their experiment, with an accuracy level of 88.7% (Kiran, 2022).

The experiment conducted by Adedayo Ogunpola et al. can serve as a potential comparison of the effect of difference in sample size and database (Ogunpola, 2024). They encompass a range of actions such as the management of missing data, encoding of variables, normalizing values of different features, and separation of datasets into training and testing groups. The sample size and database are different from the previous two experiments, retrieved from Mendeley (blue) and Kaggle (orange). The result is shown in Figure 5.

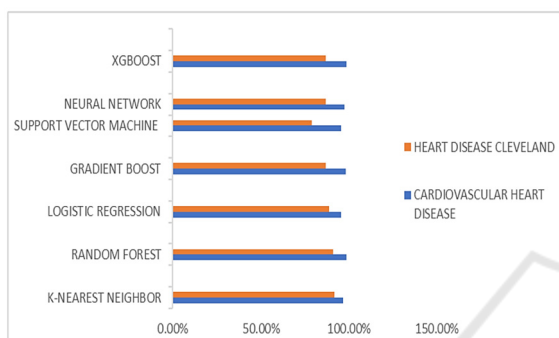


Figure 5: Graphical illustration of performance of various ML methods by Adedayo Ogunpola et al (Ogunpola, 2024).

The RF method in this experiment is 98.63% and 94.44% in precision for the two datasets from Mendeley and Cleveland respectively. This precision is seemingly higher than that obtained from the two previous experiments. What's more, the results obtained from the two databases are also different for all of the ML methods used in this report. The Mendeley database achieves a significantly higher precision than that of the Kaggle database.

From the results above, it is safe to conclude that there are already several ML methods that possess satisfactory performance after a statistical view of the outcome. However, it is also clear that the performances are influenced largely by the choice of database, attributes, sample sizes and other details on implementations of ML method. There are also several different statistical parameters describing the performance of ML methods. This provides better descriptions but also makes the comparison even harder. What's more, even the same ML method performs differently in different researches discussed above. Therefore, it is necessary to have systematic researches into the influence of choices of different attributes, sample sizes and other details.

4 CONCLUSIONS

This paper provides an extensive overview of various AI-based methods employed for predicting CVDs, particularly focusing on RF, NB, and LVQ. A notable disparity is observed among the relevant literature, highlighting the significant impact of data preprocessing methods on ML outcomes, including feature selection and handling of missing data. Additionally, sample size and database selection also play pivotal roles in influencing ML performance.

The HRFLM method presents a promising advancement over RF, while LVQ has shown superior outcomes compared to RF in one study by Saravanan Srinivasan et al., although RF performed better in other reports. Consequently, distinguishing the performance differences between these ML methods with high confidence remains challenging. Nevertheless, collectively, these ML techniques demonstrate effectiveness in facilitating CVD predictions. Despite the abundance of literature guiding the application of ML methods in CVD predictions, substantial gaps persist in data preprocessing, sample sizes, and database utilization. Hence, future research endeavors should delve deeper into exploring the impact of each factor. Additionally, it is encouraged that researchers conduct reproductions based on existing literature to further enhance understanding in this domain.

REFERENCES

- Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, 308(6943), 1552.
- Anderies, A., Tchinn, J. A. R. W., Putro, P. H., Darmawan, Y. P., & Gunawan, A. A. S. (2022). Prediction of heart disease UCI dataset using machine learning algorithms. *Engineering, Mathematics and Computer Science Journal (EMACS)*, 4(3), 87-93.
- CHERNGS. (2020). Heart Disease Cleveland UCI. <https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci>
- Damen, J. A., Hooft, L., Schuit, E., Debray, T. P., Collins, G. S., Tzoulaki, I., ... & Moons, K. G. (2016). Prediction models for cardiovascular disease risk in the general population: systematic review. *bmj*, 353.
- Gaziano, T., Reddy, K. S., Paccaud, F., Horton, S., & Chaturvedi, V. (2006). Cardiovascular disease. *Disease Control Priorities in Developing Countries*. 2nd edition.
- Gour, S., Panwar, P., Dwivedi, D., & Mali, C. (2022). A machine learning approach for heart attack prediction. In *Intelligent Sustainable Systems: Selected Papers of WorldS4 2021*, Volume 1 (pp. 741-747). Springer Singapore.

- Kiran, P., Swathi, A., Sindhu, M., & Manikanta, Y. (2022). Effective heart disease prediction using hybrid machine learning technique. *South Asian Journal of Engineering and Technology*, 12(3), 123-130.
- Mendeley Data. (2024). The Generalist Repository Ecosystem Initiative. <https://data.mendeley.com/>
- Ogunpola, A., Saeed, F., Basurra, S., Albarak, A. M., & Qasem, S. N. (2024). Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases. *Diagnostics*, 14(2), 144.
- Srinivasan, S., Gunasekaran, S., Mathivanan, S. K., M. B, B. A. M., Jayagopal, P., & Dalu, G. T. (2023). An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database. *Scientific Reports*, 13(1), 13588.

