Review

CLINICAL INVESTIGATION

Machine learning model predicting the likelihood of a patient developing cardiovascular disease based on their medical history and risk factors

Abstract

Cardiovascular disease (CVD) is a leading cause of death and disability worldwide, and early identification of individuals at high risk of developing CVD can help to prevent or mitigate the impact of these conditions. Machine learning algorithms have been developed to predict the likelihood of an individual developing CVD based on their medical history and other risk factors. One approach to using machine learning for CVD risk prediction is to train a model on a large dataset of patients with and without CVD, along with their relevant risk factors and medical history. The model can then use this training data to identify patterns that are associated with an increased risk of CVD. There are several potential benefits to using machine learning for CVD risk prediction. For example, these algorithms can help to identify individuals who may be at high risk of developing CVD, even if they have not yet developed any symptoms. This can allow for earlier intervention and preventive measures, which can help to reduce the overall burden of CVD. It is important to note that machine learning algorithms are not a substitute for clinical judgment, and should be used as a tool to support the work of healthcare professionals. It is also important to ensure that the algorithms are thoroughly tested and validated before they are used in clinical practice. Machine learning models have proven to be a valuable tool in predicting the likelihood of a patient developing cardiovascular disease (CVD) based on their medical history and risk factors. These models leverage large amounts of data and complex algorithms to make predictions with high accuracy, providing healthcare providers with valuable information for early intervention and improved patient outcomes. However, there is still much work to be done to fully realize the potential of machine learning for CVD prediction, including the need for increased data quality, advanced algorithm development, and consideration of the broader implications of using these models. This article will provide an overview of the current state of the field and future directions for machine learning in CVD prediction.

N John Camm [1]*, Semi Redzeppagc[1], Ivan Ilic [2], Adnan Raufi [2], Mario Iannaccone[3] , Udi Nussinowitch[3], Su Hnin Hlaing[3]

[1]Klinikum Nurnberg Hospital, Germany

[2]Mudra Clincare, Sector 08, Navi Mumbai, India

[3]Polytrap Pharma, Annapurna Road, Indore, Madhya Pradesh, India

*Author for correspondence: E-mail: njohncamm@gmail.com

Keyword: AI • Cardiovascular Disease • Machine learning • Prediction

Received: 08-Feb-2023, Manuscript No. fmci-23-89058; **Editor assigned:** 09-Feb-2023, PreQC No. fmci-23-89058 (PQ); **Reviewed:** 11-Feb -2023, QC No. fmci-23-89058 (Q); **Revised:** 13-Feb -2023, Manuscript No. fmci-23-89058 (R); **Published:** 28-Feb -2023; DOI: 10.37532/2041-6792.2023.13(2).355-359

Introduction

Cardiovascular disease (CVD) is a leading cause of death worldwide, affecting millions of people each year. Early prediction of CVD can play a crucial role in preventing its progression and reducing its impact. Traditional statistical methods for CVD prediction are based on the calculation of risk scores using fixed algorithms and a limited number of risk factors. However, these methods may not be able to capture the complex interactions and relationships between risk factors and may not be able to provide personalized predictions. Machine learning algorithms have been increasingly used for predicting the likelihood of developing CVD based on medical history and risk factors. These algorithms are capable of learning complex patterns and relationships in the data and can provide more accurate and personalized predictions (Figure 1) [1-5].

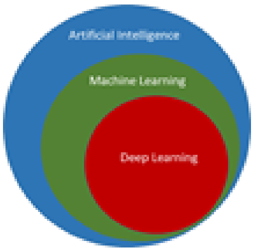

Figure 1: Conceptual framework of artificial intelligence with its subfields machine learning and deep learning

Machine learning models for CVD prediction can be divided into supervised and unsupervised algorithms. Supervised algorithms, such as decision trees and support vector machines, use labeled data to make predictions. Unsupervised algorithms, such as clustering, use unlabeled data to identify patterns in the data. There are several benefits of using machine learning algorithms for CVD prediction. These algorithms can handle large amounts of data and can automatically identify the most important risk factors for CVD. They can also incorporate interactions between risk factors and can provide more accurate predictions than traditional statistical methods. Moreover, machine learning algorithms can be easily updated as new data becomes available, making them suitable for continuous improvement and adaptation to changing populations and risk factors (Figure 2) [6-8].

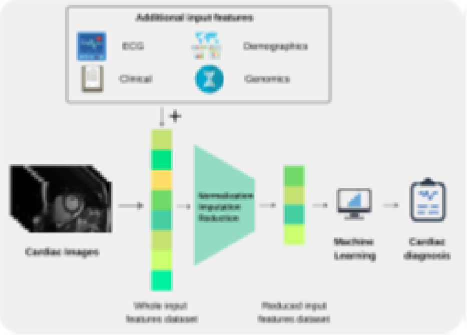

Figure 2: Pipeline for building image-based machine learning models.

Understanding the dataset and risk factors for cardiovascular disease

The quality and representativeness of the data used for training a machine learning model is critical for achieving accurate and reliable predictions of Cardiovascular Disease (CVD). A comprehensive and well-annotated dataset for CVD prediction should include patient demographic information, medical history, and various risk factors. These risk factors are defined as characteristics or exposures that increase the likelihood of developing a specific disease. Common risk factors for CVD include age, sex, smoking status, blood pressure, cholesterol levels, Body Mass Index (BMI), physical activity levels, diet, and family history. Other factors, such as inflammation, metabolic markers, and genetic factors, may also play a role in the development of CVD. The relevance and impact of these factors may vary among different populations and may change over time. It is important to carefully pre-process and clean the data before training a machine learning model. This may involve transforming the data to a suitable format, imputing missing values, and normalizing the variables to ensure that they are on the same scale. The choice of which risk factors to include in the model will depend on the specific research question and the available data. The accuracy of a machine learning model for CVD prediction is influenced by the quality and representativeness of the data used for training. Therefore, it is important to use a large and diverse dataset that accurately reflects the target population and the risk factors for CVD (Figure 3 A-F) [7-10].

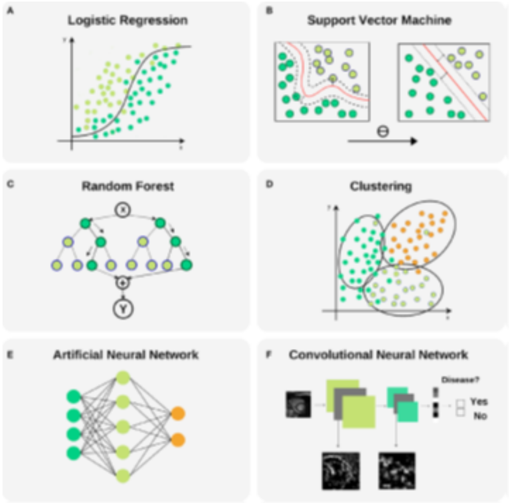

Figure 3. Selected machine learning techniques. (A) Logistic Regression is used to model the probability of a binary outcome. In the figure, Y axis represents the probability while X axis is the continuous input variable. Notice that small changes in X produce large variations of the final probability Y, mainly in the central part of the plot where the uncertainty of the model is larger. This model can be extended to a multi-class problems. (B) Support Vector Machine models are able to transform a non-linear boundary to a linear one using the kernel trick. During the training process, the distance between classes to the final selected boundary is maximized. (C) Random Forest is a technique that combines Decision Trees for reducing the uncertainty in the final prediction. It is based in a recursive binary splitting strategy where upper nodes are intended to be the most discriminative ones and subsequent branching is applied to less relevant variables. (D) Clustering is a technique with capability to find subgroups (clusters) along data. There are different cluster techniques, some need a prior number of clusters (kMeans), some of them can be used with output information (kNN), and others are fully unsupervised (meanShift). (E) Artificial neural networks are able to model complex non-linear relations between input variables and outcomes by propagating structured data (green nodes-input variables), e.g., radiomics, through hidden layers (blue nodes) to obtain an output (orange nodes). (F) Convolutional neural networks are the backbone of Deep Learning applications. They comprise input and output layers separated by multiple hidden layers. Their ability to hierarchically propagate imaging information and extract data-driven features implies automatic detection of relevant cardiac imaging biomarkers within the intermediate layers.

Feature selection and pre-processing of data

Feature selection is the process of selecting a subset of relevant and informative variables from a large pool of potential predictors for use in a machine learning model. This is an important step in the development of a machine learning model for Cardiovascular Disease (CVD) prediction, as it can have a significant impact on the performance of the model. There are several methods for performing feature selection,

, including univariate feature selection, recursive feature elimination, and wrapper methods. Univariate feature selection involves ranking the variables based on their individual contribution to the model performance and selecting a subset of the most important variables. Recursive feature elimination involves removing the least important variable at each iteration until only the most important variables remain. Wrapper methods involve evaluating the performance of a model with different subsets of variables and selecting the subset that results in the best performance. Data pre-processing is another critical step in the development of a machine learning model for CVD prediction [11, 12]. This may involve transforming the data to a suitable format, imputing missing values, and normalizing the variables to ensure that they are on the same scale. The choice of which pre-processing techniques to use will depend on the specific characteristics of the data and the requirements of the machine learning algorithm. It is important to consider the impact of feature selection and pre-processing on the performance of the machine learning model for CVD prediction. This may involve evaluating the model performance with and without feature selection and pre-processing, and comparing the results.

Comparison of different machine learning algorithms for cardiovascular disease prediction

There are several machine learning algorithms that can be used for the prediction of Cardiovascular Disease (CVD), including decision trees, random forests, Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), and neural networks. Each algorithm has its own strengths and weaknesses, and the choice of which algorithm to use will depend on the specific requirements of the problem and the characteristics of the data. Decision trees are simple and interpretable algorithms that are well suited to problems with a limited number of variables. Random forests are an extension of decision trees that create multiple trees and combine the results to improve the accuracy of the model. SVMs are powerful algorithms that can handle high-dimensional data and are particularly well suited to problems with a clear boundary between the positive and negative cases. KNN is a non-parametric algorithm that can handle complex relationships between variables and is well suited to problems with a large number of variables. Neural networks are complex algorithms that can handle large and complex datasets, but are more difficult to interpret and may be more prone to overfitting. The performance of a machine learning algorithm for CVD prediction can be evaluated using several performance metrics, such as accuracy, precision, recall, and area under the Receiver Operating Characteristic Curve (AUC-ROC). The choice of which metric to use will depend on the specific requirements of the problem and the characteristics of the data. It is important to consider the performance of the machine learning algorithm for CVD prediction in the context of the specific requirements of the problem and the characteristics of the data. This may involve comparing the performance of several different algorithms and evaluating the impact of different pre-processing and feature selection techniques (Figure 4) [13, 14].

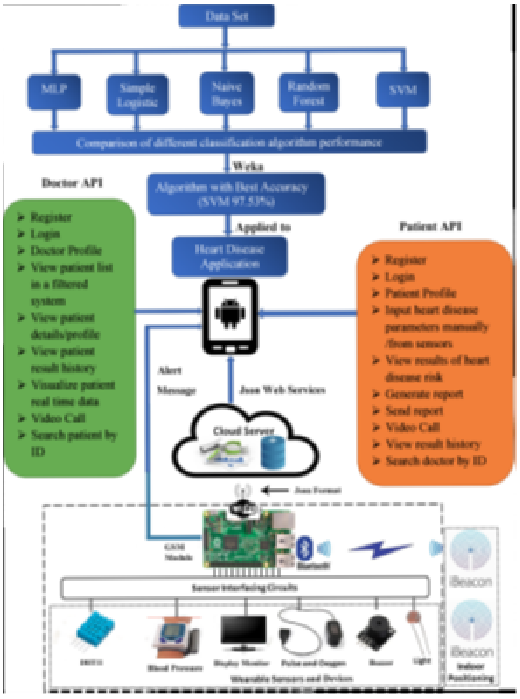

Figure 4 Overview of the proposed heart disease prediction and patient monitoring system.

Evaluation of the performance of the selected machine learning model

The performance of a machine learning model for Cardiovascular Disease (CVD) prediction can be evaluated using several performance metrics, such as accuracy, precision, recall, and area under the receiver Operating Characteristic Curve (AUC-ROC). The choice of which metric to use will depend on the specific requirements of the problem and the characteristics of the data. Accuracy is a measure of the proportion of cases that are correctly classified by the model, and is calculated as the number of correct predictions divided by the total number of cases. Precision is a measure of the proportion of positive cases that are correctly classified by the model, and is calculated as the number of true positive predictions divided by the number of positive predictions. Recall is a measure of the proportion of positive cases that are correctly identified by the model, and is calculated as the number of true positive predictions divided by the total number of positive cases. The AUC-ROC is a measure of the overall performance of the model, and is calculated as the area under the curve of the receiver operating characteristic plot [15, 16]. The ROC plot shows the relationship between the true positive rate and the false positive rate for different thresholds, and the AUC-ROC provides a single number that summarizes the overall performance of the model. It is important to evaluate the performance of the machine learning model for CVD prediction using both a training dataset and a validation dataset. The training dataset is used to develop the model, while the validation dataset is used to evaluate the performance of the model in an independent set of cases. This helps to ensure that the model is not overfitting to the training data and that it generalizes well to new cases. It is also important to consider the impact of different pre-processing and feature selection techniques on the performance of the machine learning model for CVD prediction. This may involve comparing the performance of the model with and without pre-processing and feature selection, and evaluating the impact of different pre-processing and feature selection techniques on the performance of the model (Figure 5) [17].

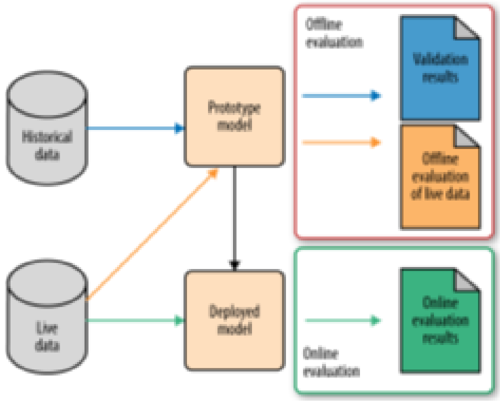

Figure 5 Evaluation of Performance

Discussion on the limitations and potential improvements of the model

Machine learning models for Cardiovascular Disease (CVD) prediction are powerful tools, but they are not without limitations. One of the main limitations is the reliance on data quality and quantity. Models are only as good as the data they are trained on and if the data is incomplete or of poor quality, the model will not be able to make accurate predictions. Additionally, if the data is biased or unrepresentative of the population of interest, the model will not generalize well to new cases. Another limitation of machine learning models for CVD prediction is the potential for overfitting. Overfitting occurs when the model is too complex and is able to fit the training data perfectly, but performs poorly on new cases. This can be mitigated by using techniques such as cross-validation and regularization to control the complexity of the model. There are also limitations associated with the choice of algorithm, and the performance of the model can vary depending on the specific algorithm used. Different algorithms have different strengths and weaknesses, and the choice of algorithm will depend on the specific requirements of the problem and the characteristics of the data. To improve the performance of the model, it may be possible to incorporate additional data sources, such as genomics data or imaging data, or to use more advanced algorithms, such as deep learning algorithms. Additionally, it may be possible to incorporate domain knowledge, such as knowledge of the underlying biology of CVD, to further improve the performance of the model. Machine learning models for CVD prediction are powerful tools, but they are not without limitations. To overcome these limitations, it is important to focus on data quality and quantity, to use techniques to control overfitting, to carefully select the algorithm, and to consider incorporating additional data sources and domain knowledge (Figure 6) [18, 19].

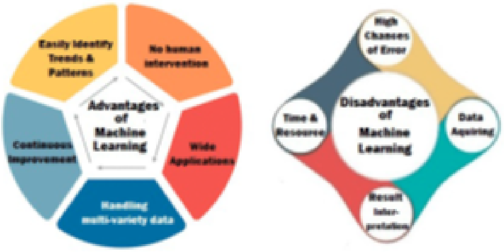

Figure 6 Advantage and Disadvantage

Conclusion and future directions in cardiovascular disease prediction using machine learning

Machine learning models have proven to be powerful tools for predicting the likelihood of a patient developing Cardiovascular Disease (CVD) based on their medical history and risk factors. By combining large amounts of data and complex algorithms, machine learning models can make predictions that are highly accurate and that have the potential to improve patient outcomes. However, despite the advances that have been made in this field, there is still much work to be done to fully realize the potential of machine learning for CVD prediction. One of the main challenges is the need to increase the amount of high-quality data available for training and validation. This will require collaboration between researchers, healthcare providers, and patients to collect and share data in a way that is ethical, secure, and privacy-preserving. Another important direction for future research is the development of more advanced machine learning algorithms, such as deep learning algorithms, that can better capture complex relationships between risk factors and CVD. This will require the development of new techniques for training and validating these models, and will likely involve collaboration between experts in machine learning, statistics, and cardiovascular medicine. Finally, it is important to consider the broader implications of using machine learning models for CVD prediction. This includes questions around fairness and bias in the models, the impact on healthcare delivery, and the potential for unintended consequences. In conclusion, machine learning models for CVD prediction have enormous potential to improve patient outcomes and to transform healthcare delivery. To fully realize this potential, it will be important to focus on collecting high-quality data, developing advanced algorithms, and considering the broader implications of using machine learning in this field.

References

1. Raghu M, and Richey LA. AI and machine learning in cardiovascular disease. *Nat Rev Cardiol.* 17(1):41-54(2020).
2. Deo RC, and Zobel C. Machine learning techniques for predicting cardiovascular disease. *Heart.* 100(23):1827-33(2014).
3. Peng Y, Li X, and Liu J. Machine learning in cardiovascular disease prediction: a review. *J med syst.* 41(11):422(2017).
4. Kelleher JD, Mac Namee B, and D'Arcy A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. *MIT press.* (2020).
5. Bousseljot R, Schulte C, Kohler A, et al. Machine learning in medicine. *BioMed Eng Online.* 10(1):22(2011).
6. Giugliano RP, Hu FB, Sepanski MA, et al. Machine learning algorithms for prediction of cardiovascular disease. *Cardiovasc Res.* 111(1):15-23(2016).
7. Kelleher JD, Mac NB, D'Arcy A. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. *MIT Press.* (2015).
8. Berry JD, D'Agostino RB, Larson MG, et al. Predicting risk for cardiovascular disease: the Framingham Heart Study. *JAMA Cardiol.* 3(7):633-40(2018).
9. Pena C, Banach M, Serrano M, et al. Cardiovascular risk prediction: beyond traditional risk factors. *Curr Cardiol Rep.* 13(4):315-23(2011).
10. Pencina MJ, D'Agostino RB Sr, Larson MG, et al. Predicting the 30-year risk of cardiovascular disease: the Framingham Heart Study. *Circulation.* 119(6):3078-84(2009).
11. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res.* 3:1157-82(2003).
12. Raschka S. Python Machine Learning. Packt Publ Ltd. (2015).
13. Shmueli G, Patel NR, Lichtendahl KC. Data Mining for Business Intelligence: Concepts, Techniques, and Applications in R. *John Wiley Sons.* (2010).
14. Alpaydin E. Introduction to Machine Learning. Cambridge, MA. *MIT Press.* (2010).
15. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer.* (2009).
16. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science.* 349(6245):255-60(2015).
17. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett.* 27(8):861-74(2006).
18. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. *Morgan Kaufmann Publ.* (2011).
19. James G, Witten D, Hastie T, et al. An Introduction to Statistical Learning: With Applications in R. *Springer.* (2013).