DOI: 10.55415/deep-2023-0019.v1

Machine learning methods in real-world studies of cardiovascular disease

Jiawei Zhou[1#], Dongfang You[1#], Jianling Bai[1], Xin Chen[1], Yaqian Wu[1], Zhongtian Wang[1], Yingdan Tang[1], Yang Zhao[1*], Guoshuang Feng [2,3*]

Affiliations

[1] Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu, China, 211166

[2] Big Data Center, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China, 100045

[3] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University & Capital Medical University, Beijing, China, 100083

[#]These authors contributed equally to this work.

***Correspondence:**

Guoshuang Feng, PhD, E-mail: glxfgsh@163.com

Yang Zhao, PhD, E-mail: yzhao@njmu.edu.cn

Total words: 4608,

Abstract: 224

DOI: 10.55415/deep-2023-0019.v1

Machine learning methods in real-world studies of cardiovascular disease

Abstract

Objective: Cardiovascular disease (CVD) is one of the leading causes of death worldwide and multiple questions urgently need answering, especially in risk identification and prognosis prediction. Real-world study (RWS), with huge numbers of observations, is an important data basis for CVD research, but it is constrained by high dimensionality, missing, and unstructured data. Machine learning (ML) methods, including a variety of supervised and unsupervised algorithms, are useful for data governance and effective for high dimensional data analysis and imputation in the real-world study. This study reviewed the theory, strength, limitation, and application of several popular ML methods in the CVD field as a reference for further application.

Methods: This study introduced the origin, purpose, theory, superiorities, limitations, and applications of multiple popular ML algorithms, including hierarchical and k-means clustering, principal component analysis, random forest, support vector machine, and neural networks. An example using the Systolic Blood Pressure Intervention Trial (SPRINT) data was performed with the random forest to demonstrate the process and main results of ML application in CVD.

DOI: 10.55415/deep-2023-0019.v1

Conclusion: ML methods are effective tools to produce real-world evidence to support clinical decisions and meet clinical needs. This review explains the principles of multiple ML methods in an easy-to-understand language and could be a reference for further application. Future research is warranted for accurate ensemble learning methods and wide application in the medical field.

Keywords: Cardiovascular disease, Machine learning, Real-world study.

DOI: 10.55415/deep-2023-0019.v1

Introduction

Cardiovascular disease (CVD) is the leading cause of death worldwide, killing 17.9 million people each year[1]. A large number of randomized clinical trials (RCTs) are conducted to evaluate the efficacy and safety of CVD treatment interventions, as well as the primary and secondary prevention of CVDs, including drugs like statins[2, 3] and polypills[4], dietary intakes (such as Mediterranean diet and dietary supplements[5-7]), behaviors[8] or lifestyles (such as weight loss[9]). The importance of RCTs with large sample sizes has been well-recognized. Although RCTs can generate the most credible and highest-level evidence to assess the prevention and treatment effects of CVDs, their applications are limited by cost, duration, lack of generalizability, ethical concerns, and technical feasibility[10, 11].

Real-world study (RWS) has been recognized as an appealing alternative to RCT in recent years[10, 12]. Real-world data (RWD), collected in routine health care from different sources, include electronic health records (EHRs), registry cohorts, health claims, and records from home-use settings or mobile devices, etc. [13]. RWS utilizes RWD to generate different levels of real-world evidence (RWE) [14]. While analysis of confirmatory RCT relies on traditional statistical methods, there is growing interest in the application of ML to address the challenges posed by RWD analysis, such as high-dimensional, complex and unknown data patterns, and the rapid growth of data volume[15, 16].

DOI: 10.55415/deep-2023-0019.v1

ML is a family of methods focusing on classification and prediction[17]. Combined with increasing computational capacity, ML methods have ushered in a new era of medical research analysis (figure 1). There are numerous successful applications of ML methods in data governance, risk factor identification, and outcome prediction based on RWD. Being excelled in discovering potential influencing factors and non-linear relationships, ML methods would make the analysis of RWS get a new pulse with high efficiency.

A recent retrospective analysis of transversal RWS in more than 11,000 patients over 65 years old was performed using principal component analysis (PCA), clustering, synthetic minority oversampling technique, and logistic regression to diagnose cardiac amyloidosis, a rare disease with poor diagnosis resulted in untimely treatment[18]. These analyses with high dimensionality, low prevalence, and missing data in EHRs filled by structured and unstructured records relied on the processing and pipelines of data governance and analysis by ML algorithms to transform the investigation of cardiac amyloidosis into a new pattern that met patient needs. Furthermore, among 13,602 patients with heart failure, multiple ML methods, including support vector machine (SVM), artificial neural network (NN), random forest (RF), and extreme gradient boosting models, were displayed for prognosis prediction with the area under the receiver operating characteristic curves (AUC) higher than 0.85,

DOI: 10.55415/deep-2023-0019.v1

which met the clinical demand[19]. Thus, RWS and ML are critical tools to fill knowledge gaps and meet the medical needs for the research of CVDs, one of the most complex diseases.

This study reviewed the ML methods commonly used in CVD-related studies. Although not exhaustive, this review could be a reference for application in RWS of CVDs. This review is organized as the following. The principles and algorithms of several popular ML methods will be first introduced. Then, an example based on the analysis of the Systolic Blood Pressure Intervention Trial (SPRINT) data is provided to illustrate the basic procedures of applying ML methods. Finally, the advantages and limitations of ML methods are discussed.

Machine learning Methods

Machine learning (ML) algorithms derived from the 1950s[20]. Benefitting from novel learning algorithmic boom, vastly improved computational power, and enormous and still-increasing RWD[21], data-intensive ML is able to mine clinical data in large volumes and/or across large time scales. ML methods could be categorized into supervised and unsupervised learning methods depending on whether an outcome variable is specified (labeled/unlabeled).

Unsupervised learning

DOI: 10.55415/deep-2023-0019.v1

The main task of unsupervised ML is to explore the hidden data pattern and group unlabeled data into sub-population by clustering and/or dimensionality reduction with feature/variable selection. Because unsupervised learning methods can identify the underlying data structure without the need for human intervention, they are quite suitable for exploratory analysis[22].

Clustering Analysis

Clustering analysis is not a specific algorithm, but a general task to set objects into two or multiple sub-groups. The first definition of cluster analysis was originally proposed by Driver and Kroeber in 1932[23]. Clustering analysis aims to find distinct groups or “clusters” of individuals or characters based on the distance among them.

There are mainly two types of clustering: sample clustering and variable clustering[24]. Variable clustering can use similarity metrics such as correlation coefficients to find similar variables. When two variables are found in a cluster, one of them can further be considered a “surrogate” for the other. Sample clustering procedures are used to classify individuals into different subgroups based on the distance between individuals.

DOI: 10.55415/deep-2023-0019.v1

For clustering algorithm, the definition of the distance is used as the similarity measure of data points or samples. Some commonly used distance measures are displayed in Table 1. We denote x, y are samples with N features/variables. x_i, y_i are the numerical value of i^{th} feature. In most cases, data should be normalized before clustering to eliminate the effects of heterogeneity and variation. It is easy to see that Manhattan Distance, Euclidean distance, and Chebyshev Distance are special cases of Minkowski distances (p=1,2 and ∞, respectively) [25-28].

Hierarchical clustering[29], the earliest clustering methods used by biologists and social scientists[30], aims to create clusters in a hierarchical tree-like structure. The algorithm defines distance for each pair of data points, selects the closest data pair, groups them together, and updates the representation value of the data pair with the mean or median at each step. These steps are repeated to include all the samples in the hierarchical cluster tree.

One of the most popular clustering methods is k-means clustering[31]. Firstly, k, the number of clusters, should be determined a priori. K cluster centers are selected randomly. Then, each data point would be categorized into its closest center, which is represented by the mean or median of all data points assigned. Finally, repetition of the process of categorization and update of the center presentation would be performed until it converges to k optimal

DOI: 10.55415/deep-2023-0019.v1

clusters, each of which would not be changed or changed in a small range. The later k-means++ algorithm optimizes cluster center selection by selecting the first center randomly and taking the distance as the probability to select the other k-1 cluster centers[32].

As CVDs are multicausal diseases, the interactions among different risk factors are complex. The clustering could be used for combination of risk factors. As shown in Table 2, Bel-Serrat et al.[33] investigated association of lifestyle behaviors with CVD risk factors. Hierarchical and k-means clustering were performed for measurements, including dietary consumption, physical activity performances, and video viewing in children. Clusters were converted into dummy variables and described characteristics. This research found lower levels of video viewing and consumption of sugar-sweetened beverages were associated with healthier cardiovascular outcomes than diets rich in fruits and vegetables or physical exercise.

All these methods have advantages and disadvantages, and should be selected according to the properties of the application data. K-means clustering is intuitive, straightforward, and easy to handle. However, the value of k is pre-specified by users, which may depend on the visualization. The clustering results are strongly influenced by k, which has no objective optimal value. Modifications were proposed for the optimal selection of k[31]. Instead of specifying a single value of k, a set of values might be considered in application. For

DOI: 10.55415/deep-2023-0019.v1

hierarchical clustering, a noniterative, single-pass greedy algorithm, the measure of distance depends on specific data and should be chosen carefully. Hierarchical clustering has intuitive tree-like structure output but lacks best solution for cluster segmentation, so it can only depend on professional knowledge.

Dimensionality Reduction

Informally, the curse of dimensionality induced decreased computational power, high variance, or overfitting, with the exponential increase of features/variables [34, 35]. Although higher dimensions theoretically include more information, it hardly helps due to noise, redundancy, and sparsity in practice. Avoiding the curse of dimensionality, the dimensional reduction could be considered for initial exploratory analyses, with common approaches PCA (principal component analysis)[36] and SVD (singular value decomposition)[37]. We will mainly focus on the PCA algorithm.

PCA[36] is one of the most popular statistical algorithm to reduce dimensionality. PCA was originally invented by Pearson in 1901 and further developed by Hotelling to its present form[38]. The purpose of PCA is to decrease dimensionality while minimizing information loss. The aim is accomplished by linear transformation and combining original variables into a new coordinate system.

DOI: 10.55415/deep-2023-0019.v1

We denote a feature space with *N* features ($\boldsymbol{X_N}$,variables). Linear combination of features with maximum variance are sought, which presented as $\sum_{i=1}^{N} a_i X_i$ with $\boldsymbol{a}$ as a vector of constants $\boldsymbol{a = (a_1, \ldots, a_N)}$. The variance of each linear combination could be written as $Var\,(\sum_{i=1}^{N} a_i X_i) = \boldsymbol{a'Sa}$, with **S** from covariance or correlation matrix and $'$ denoting transpose. With common restriction of $\boldsymbol{a}$ as a unit vector, solution of the maximum variance searching could be reduced to maximize the equation $\boldsymbol{a'Sa - \lambda(a'a - 1)}$, then with derivation process, presenting as $\boldsymbol{Sa - \lambda a = 0}$. $\boldsymbol{a}$ is the eigenvector; λ is the eigenvalue; and linear combination is the corresponding principal component. The first principal component has greatest variance, and the second principal component follows. This greatest linear combination searching process lasts until the *N* principal component. To reduce dimensions, first *n* principal components are selected for the following analysis with some information loss. Visualization of cumulative information (percentage of explained variances) with ordered principal components could be used for selection of *n.* Based on original and widely used PCA, novel methods for further dimensional reduction have been proposed, such as kernel PCA[39], t-Distributed Stochastic Neighbor Embedding[40], and nonmetric multidimensional scaling[34, 41, 42].

DOI: 10.55415/deep-2023-0019.v1

PCA is intuitive, easy to apply and not limited by the number of variables. However, PCA has its own demerits. Underlying assumption of PCA is the relationship between variables is linear. In the case of non-linearity, PCA may produce inaccurate results. Additionally, PCA can efficiently reduce the dimensions for related features but does not perform well for uncorrelated situations.

Large datasets are increasingly being utilized in the exploration of CVD domain. PCA is one of the optimal choices for dimensionality reduction. Peterson et al.[43] used PCA to determine a continuous metabolic syndrome score (MetScore) as a cardiometabolic risk pattern with waist circumference, fasting glucose, systolic blood pressure, triglycerides, and glucose. They further examined association between MetScore and age, body mass index, cardiorespiratory fitness (CRF), physical activity (PA), and parental factors. They claimed independent contribution of CRF, PA, and family-oriented healthy lifestyles to improve the health of 6^{th} graders.

Unsupervised ML will be suitable to identify subgroup of population with patient profile, which may be utilized in precision medicine. The more detailed the information, the more accurate the prediction. However, for RWD containing thousands of measures of complex information, the visualization, analysis, and interpretability of data are challenging due to ‘the

DOI: 10.55415/deep-2023-0019.v1

curse of dimensionality'. Application of clustering or PCA would provide partial solutions, and some algorithms have been proposed for prediction with high dimensions[44, 45].

Supervised learning

In contrast to unsupervised algorithms, supervised algorithms predicted an outcome class (probability) or value with a pre-specified label. Supervised learning algorithms are trained with input datasets to detect the underlying patterns and relationships with labels (supervisory signal).

Random forest

RF (Random forest), proposed by Breiman in 2001[46], is a preferable classification and regression algorithm in recent years[44]. Basically, RF is a combination of multiple decision trees and aggregating predictions by averaging or voting. The growth steps of trees are as follows: firstly, for the dataset with M samples and N variables, a dataset with M observations are randomly sampled with replacement (bootstrap sample) from original training dataset (bagging step). The number of the remaining sample, called 'out-of-bag' (OOB) sample, is approximately equal to a third of M. Then, *n*, with default value as the square root of N, is pre-specified for each node. For separation, *n* variables are randomly selected from N input covariates, and best split is performed for maximum 'purity' with these *n* features. Purity

DOI: 10.55415/deep-2023-0019.v1

represents average differences and proportions of continuous and categorical predictive separation variables, respectively, which could be presented by entropy or Gini index in random forests[47]. This separation step is iterated for each subset until there are too few samples in final subset. Generally, this iteration will obtain an oversized tree with the overfitted phenomenon, which presented as a small bias but a large variance. To overcome this situation, cross-validation could be used for pruning.

RF is a constitutive supervised method, combined multiple decision trees for prediction. There is no pruning for decision trees, which means each tree is grown to the largest extent. For prediction step, classifying a new sample with inputting variables to the forests can obtain multiple outcomes. The model chooses classification with most votes in the RF as final prediction of this new sample. Further extensions to original RF ranged from weighted forests with tree-level weight for more accurate prediction[46, 48], online forests with streaming input dataset[49], random survival forests incorporating survival endpoints[50], clustering forests in the context of unsupervised classification[51], ranking forests for ranking problems[52], to the forests with correcting confounding bias for removing spurious association[53].

DOI: 10.55415/deep-2023-0019.v1

Aryal et al.[54] investigated the gut microbiome-based diagnostic screening of CVD using different ML methods. With top 500 high-variance features of operational bacterial taxonomic units, RF could achieve prediction of CVD with an AUC as 0.65[54]. Other successful applications of RF algorithms in the CVD area include CVD prediction with all collected baseline variables and top-20 predictors selection[55], descriptors identification of coronary CT angiography imaging and fractional flow reserve features for ischemia-related lesion prediction[56], and mortality prediction in aortic stenosis patients with cardiovascular magnetic resonance measures[57].

RF algorithm has advantages. Firstly, RF has the merits of high accuracy and efficiency. Secondly, it can handle tens of thousands of input variables, even much larger than the number of observations, without any variable deletion, which is an appealing characteristic in RWS. Additionally, it can accommodate well in the scenario with interactions among predictive or prognostic variables. Furthermore, it returns measures of variable importance, which is helpful for clinicians to identify important predictors for disease susceptibility or prognostics. However, even after removing confounding effects, the variable importance itself did not indicate causality among outcome and variables, which needs further mechanistic research. Additionally, standard RF classifier has great performance on balanced data, but may perform poorly in the case of extremely unbalanced classes. For the common unbalanced

DOI: 10.55415/deep-2023-0019.v1

data in the cardiovascular domain, an over-sampled version is often used to fit the model, resulting in insufficient variability in the minority class and poor performance in a giant range prediction. Weighted RF or AdaBoost should be considered for application[58, 59].

Support Vector Machine

SVM (support vector machine) was initially proposed by Boser, Guyon, and Vapnik in 1992[60]. The purpose is to search for optimal boundary (decision surface/ hyperplane) in a multi-dimensional space that completely classifies data points with largest gap (distance) between borderline features (support vectors).

Taking the two-dimensional data in figure 2a as an example, both CVD patients and normal people have feature data of M and N. SVM algorithm find the line completely distinguishes the participants with largest distance between borderline non-CVD and CVD patients. Obviously, patients are not always linearly separable in real-world data (figure 2b). SVM algorithm maps the data into a much higher dimensional space (high dimensional features) where the separatable decision surface (hyperplane) could be selected (figure 2c). Kernel functions are used for dimensional map:

$$k(x_i, x_j) = < \phi(x_i)\phi(x_j) >.$$

DOI: 10.55415/deep-2023-0019.v1

x_i is the pre-mapping data for patient i; ϕ is mapping function; k is the kernel function. A kernel function calculated in input feature space corresponds to a dot product in some feature space if and only if it is a symmetric positive definite function. The map can be achieved without an explicit map function, but rather a detailed kernel function, due to the distance from points to plane transformed and simplified into the dot product instead the map function. One commonly used kernel is gaussian kernel function:

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right).$$

Detailed mathematical principles and proofs can be found elsewhere[45, 61]. Additionally, SVM algorithm introduced slack variable δ for the soft margin of decision surface. The slack variable δ relaxes restrictions of linearly separable to avoid extreme inseparability. Later extensions of SVM include multi-classes SVM[62, 63], transductive SVM[64], and Bayesian SVM[65, 66].

SVM [45] for classification and support vector regression for continuous data are widely used in computational biomedicines. SVM had best performance with the highest accuracy compared with logistic regression and NNs in predicting the prevalence of CVD using health-related data measured by smartwatch from the Korea National Health and Nutrition Examination Survey[67]. Additionally, Petrazzini et al. applied an ML framework, including

DOI: 10.55415/deep-2023-0019.v1

PCA, RF, and SVM, using features from EHR to increase prediction and reclassification for coronary artery disease[68].

Just like other ML methods, SVM has merits and demerits. Firstly, it can be computed with very many variables and small samples with robust prediction, which is a satisfactory advantage for high-dimensional RWD. Additionally, SVM was designed based on sophisticated mathematical principles, which could avoid overfitting. However, an important demerit of SVM is the subjective choice of kernel function, which often depends on repeated tries. Distance maximization search obtained optimal results, but at the expense of huge computing requirements. Additionally, the hyperplane was determined by the sample closest to the borderline, which would generate a perfect classification with the biggest margin or be affected by overlapping outliers from different classes in an infinite loop.

Neural Networks

The spring-up of ML algorithms could be traced back to the induction of artificial neural networks (NNs) by McCulloch and Pitts in 1943[69, 70]. NNs get their name and structure from imitation of the biological neuron of human brain transmitting signals from one neuron to another, to model complex patterns and prediction problems. There are different layers in the NNs, including an input layer, one or more hidden layers, and an output layer. For

DOI: 10.55415/deep-2023-0019.v1

example, in figure 3, the x_1, x_2 are input feature variables, the a_1, a_2, z_1, z_2 are hidden nodes, and $\breve{y}$ is the output outcome.

In this 3-layer NN, the potential formulation follows:

$$a_1 = w_{11}x_1 + w_{12}x_2 + b_1$$

$$a_2 = w_{21}x_1 + w_{22}x_2 + b_2$$

$$z_1 = \varphi(a_1)$$

$$z_2 = \varphi(a_2)$$

$$\breve{y} = w_1 z_1 + w_2 z_2 + b_3$$

For the notion, $w_{11}, w_{12}, b_1, w_{21}, w_{22}, b_2, w_1, w_2, b_3$ are parameters to be estimated. $\varphi(\cdot)$ is a non-linear step function, which can be sigmoid function or hyperbolic tangent function with S shape.

$$sigmoid\ function: \varphi(x) = \frac{1}{1+e^{-x}}$$

$$hyperbolic\ tangent\ function: \varphi(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

There are two core parts in the NNs. Firstly, the users should decide the number of layers and nodes in advance but without an objective or standard answer to this question. These numbers are often determined by experience. Some scholars thought this is one of the serious demerits of the NNs. Secondly, training dataset with m data points is input into the algorithm to estimate parameters with the cost (or loss) function:

DOI: 10.55415/deep-2023-0019.v1

$$L(\boldsymbol{w},\boldsymbol{b}) = argmin \sum_{i=1}^{m} (y - \breve{y})^2.$$

Additionally, a regularization term can be added to the cost function as a penalized part of model simplification. The updated cost function is:

$$E(\boldsymbol{w},\boldsymbol{b}) = L(\boldsymbol{w},\boldsymbol{b}) + \frac{\lambda}{2}||\boldsymbol{w}||^2.$$

λ is the weight decay parameter. The gradient descent method could be used for the parameter estimation. As the foundation of deep learning, later multiple extensions include but are not limited to feed-forward NNs[71], recurrent NNs (RNN)[72], modular NNs[73], deep NNs[74], and convolutional NNs (CNN)[75, 76], with applications ranging from medicine to multiple domains around our daily life.

Recently, derivative algorithms based on NN have been used in cardiac mechanics[77], genetic variants[78], and electrocardiographic diagnosis[79]. Narula et al.[80] investigated application of ML methods, SVM, RF, and NNs, to identify physiological and pathological patterns of hypertrophic remodeling. Expert-annotated speckle-tracking echocardiographic datasets were utilized to develop a machine-learning-based automated system for the interpretation of echocardiographic images.

As one of the most important ML algorithms, NNs also has its advantages and disadvantages. It could recognize and learn the non-linear and complex relationship for

DOI: 10.55415/deep-2023-0019.v1

modeling and prediction with many achievements in recognition and prediction. It has the capacity to handle non-structural data and output multiple classifications. However, NNs generally require much more data for learning than other algorithms mentioned above, which leads to limited application. “Black box” property of NNs is obvious with multiple unexplainable parameters, making it difficult for parameter modifications and interpretation. Meanwhile, the gradient descent method is prone to local minima during model training.

Application of Machine Learning Methods in Real Data

Systolic Blood Pressure Intervention Trial (SPRINT) data was utilized to perform analysis with ML algorithms in RWD for application demonstration. SPRINT was a randomized, controlled, open-label trial for specifying appropriate targets for systolic blood pressure (SBP) to reduce cardiovascular morbidity and mortality among people without diabetes. Details of this trial, which included 9,361 people with SBP of 130 mm Hg or higher, were described previously[81].

Although SPRINT is designed to be a clinical trial, its characteristics of the long-term follow-up and diverse treatment regimen make it as a good example for our introduction of the application of ML methods in RWS analysis. All information collected at baseline was applied to predict composite CVD outcomes, including myocardial infarction, stroke, heart

DOI: 10.55415/deep-2023-0019.v1

failure, non-MI acute coronary syndrome, or CVD death. Based on the updated 2020 SPRINT data, incidence of composite outcomes is 7.8%, with 726 events. Considering imbalanced distribution of the outcomes, we randomly sampled 1,452 observations from the non-event population, resulting in an example with 2,178 participants. Baseline information, including from demographics, medical history, clinical status, anthropometry, laboratory, and ECG data, was merged with more than 120 variables (Supplementary Table 1). The example was divided into training and test dataset with 70% and 30% participants. The supervised methods, RF, SVM, and NNs, were performed. R software (version 4.1.2) was used for analysis (Supplementary Methods).

The RF prediction showed an accuracy of 0.71. The top-10 variables and ROC curve were displayed in figure 4. SVM and NNs prediction was further performed, achieving an accuracy of 0.70 and 0.68, respectively. Here, it should be noticed that we present this example just as an illustration of how to apply the ML methods in CVD research. Machine learning methods are mainly focused on data mining with continuous and categorical data. In the field of cardiovascular disease, survival outcomes are common, which could be transferred to categorical data, including mortality and morbidity. Meanwhile, restricted mean survival time and pseudo-survival methods could be utilized as outcome calculations for further application with machine learning methods. Some algorithms are also extended to survival data, such as random survival forests. In this application, we tried random survival

DOI: 10.55415/deep-2023-0019.v1

forests for SPRINT data with AUC as 0.71 at 4 years and 0.63 at 5 years after treatment. Additionally, both internal and external validation are critical for RWE in CVDs.

Discussion

In this paper, we conducted a comprehensive summary of recent ML algorithms, introduced principles behind these methods, cited relative applications of CVDs, and displayed an application of ML in CVDs. We believe more ML methods would be applied using big databases to provide RWE in CVDs domain to support clinical diagnosis, treatment selection, and prognosis prediction. Future research should routinely focus on ensemble learning of ML methods and be widely applicated in medical domains.

RWE (real-world evidence), is a hot spot in medical research, comprising information produced from RWS (real-world study) [14]. Causal inference, especially causal model is an important component of RWE [82]. However, although there are quite a lot of examples of RWS in the background of regulatory decisions, RWS by its own is a class of evidence-generation study type, including observational studies and pCT (pragmatic clinical trial). The generation and application of RWE commonly included multiple parts, comparison of effectiveness and safety for regulatory decisions, market assessment, health economic evaluation, clinical trials design, and predictive and prognostic factors identification for

DOI: 10.55415/deep-2023-0019.v1

disease exploration and improved healthcare delivery, etc. The generation of high-level evidence based on RWS relies on not only scientific design, but also appropriate analysis methods [82]. Machine learning methods are useful and effective tools for the generation of RWE [83]. Supervised methods could be applied for risk prediction and predictive and prognostic factors identification; unsupervised algorithms could be utilized for dimensional reduction and classification.

The traditional statistical methods applied expert opinions or rules to the collected data for analysis. ML algorithms learn patterns from data and feed them back to search for potential relation, which are then further validated by other research with high-level evidence, finally guiding the clinic. The difference lights on the ML algorithms' expectation to solve problem beyond human capability in complex diseases, especially in CVDs with ECG and medical images. Traditional imaging diagnosis could only rely on the professionalism and experience of doctors. However, ML could handle image recognition effectively via CNN. The utilization of ML algorithms could help doctors spot the focus and analyze image data with clinical records. Additionally, RWS has the characteristic of high dimensionality, missing and unstructured, which is still unresolved for traditional statistical methods. ML methods could be useful to address these issues partially or comprehensively. Most methods

DOI: 10.55415/deep-2023-0019.v1

in this review could be applied with high-dimensional data. As for missing, ini with ML algorithms can be reviewed in other papers [84, 85].

In the past 5 years, there were more than 4,000 published papers on ML application of CVDs in PubMed (figure 1), almost were related to the diagnosis, classification, and prognosis prediction of CVDs[86]. The introduction of ML into CVDs facilitates the extraction of features from EHRs, medical images, and laboratory tests[87]. ML methods could be applied to rare or complex disease for timely treatment with better prognosis to fill the medical gap. More than half of the present applications focus on atherosclerosis, heart failure, hypertension, and other cardiac risk factors[86]. Other areas of CVDs require further research with ML methods accepted and utilized in a wider range.

In the future, the intelligence of ML in data preparation and analysis should be optimized for better performance and interpretability. ML has limitations, but there are opportunities for exploration. Firstly, there is still room for improvement in accuracy, robustness, and interpretability of these methods for different applications. Ensemble learning, with sophisticated mathematical theories, aimed to build a unified framework to integrate data fusion, modeling, and mining [88]. It combined several models via voting in an adaptive way to improve machine learning results and produced better predictive performance when

DOI: 10.55415/deep-2023-0019.v1

compared to a single model, especially dealing with imbalanced and noisy data [89]. More ensemble methods are warranted for better performance. Secondly, most ML methods are black boxes with poor interpretability. This property will limit its application. The classification based on ML without proper interpretation is difficult for clinicians and patients to accept. Additionally, identification of risk factors and the estimation of treatment effect play important roles. The “black-box” property of ML limits its application in these aspects. Furthermore, most ML methods are applicable independently. Future systems should be capable of working collaboratively[21] with massive different joint data to explore potential correlation and causality.

Future applications should also consider deep learning, which is part of the broader family of ML. Deep learning is based on NNs with multiple layers and presentation learning with higher-level features extracting[90]. Deep learning methods had a great impact and dramatically improved accuracy level in digital processing of images, video, speech, and audio with CNN and RNN[91]. These methods would make great advances in the medical domain, including natural language processing (NLP) using multiple algorithms to analyze free text and generate structured presentations for EHRs[92], and extracting features from medical images, including PET(positron emission tomography)/CT[93], EEG (Electroencephalography)[94], and ECG (Electrocardiography)[95] with CNN[96].

DOI: 10.55415/deep-2023-0019.v1

Conclusion

In this paper, we conducted a comprehensive review of ML algorithms, including supervised and unsupervised methods. This tutorial could be served as a reference for ML application in CVDs. In summary, ML algorithms bring new strengths to data mining in RWD, but there are still some limitations. Future work is warranted in both methodology development and CVD application.

Acknowledgements

This study was funded by the National Natural Science Foundation of China (Project No. 82173620 to Y.Z., 82204156 to D.Y.). This study was also funded by the Priority Academic Program Development of Jiangsu Higher Education Institution (PAPD).

Conflicts of Interest

None.

DOI: 10.55415/deep-2023-0019.v1

Reference

1. WHO.int [website on the Internet]. Cardiovascular diseases; [updated 2019 Jun 11; cited 2022 Nov 1]. Available from: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.
2. Taylor F, Huffman MD, Macedo AF, Moore TH, Burke M, Davey Smith G, et al. Statins for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev 2013(1):CD004816.
3. Chou R, Dana T, Blazina I, Daeges M, Jeanne TL. Statins for Prevention of Cardiovascular Disease in Adults: Evidence Report and Systematic Review for the US Preventive Services Task Force. JAMA 2016;316(19):2008-2024.
4. Chow CK, Meng Q. Polypills for primary prevention of cardiovascular disease. Nat Rev Cardiol 2019;16(10):602-611.
5. Bhatt DL, Steg PG, Miller M, Brinton EA, Jacobson TA, Ketchum SB, et al. Cardiovascular Risk Reduction with Icosapent Ethyl for Hypertriglyceridemia. N Engl J Med 2019;380(1):11-22.
6. Group ASC, Bowman L, Mafham M, Wallendszus K, Stevens W, Buck G, et al. Effects of n-3 Fatty Acid Supplements in Diabetes Mellitus. N Engl J Med 2018;379(16):1540-1550.
7. Howard BV, Van Horn L, Hsia J, Manson JE, Stefanick ML, Wassertheil-Smoller S, et al. Low-fat dietary pattern and risk of cardiovascular disease: the Women's Health Initiative Randomized Controlled Dietary Modification Trial. JAMA 2006;295(6):655-666.
8. Nepper MJ, McAtee JR, Wheeler L, Chai W. Mobile Phone Text Message Intervention on Diabetes Self-Care Activities, Cardiovascular Disease Risk Awareness, and Food Choices among Type 2 Diabetes Patients. Nutrients 2019;11(6).
9. Look ARG, Pi-Sunyer X, Blackburn G, Brancati FL, Bray GA, Bright R, et al. Reduction in weight and cardiovascular disease risk factors in individuals with type 2 diabetes: one-year results of the look AHEAD trial. Diabetes Care 2007;30(6):1374-1383.
10. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther 2017;102(6):924-933.
11. Chen D. Real-world studies: bridging the gap between trial-assessed efficacy and routine care. J Biomed Res 2022;36(3):147-154.
12. McNair D, Lumpkin M, Kern S, Hartman D. Use of RWE to Inform Regulatory, Public Health Policy, and Intervention Priorities for the Developing World. Clin Pharmacol Ther 2022;111(1):44-51.
13. Real-Word Evidence [website on the Internet]. Real-world data (RWD) and real-world evidence (RWE) are playing an increasing role in health care decisions; [updated 2022 Dec 12; cited 2023 Jan 11]. Available from: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.
14. Schad F, Thronicke A. Real-World Evidence—Current Developments and Perspectives. International Journal of Environmental Research and Public Health 2022;19(16):10159.
15. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci 2021;2(3):160.
16. Brnabic A, Hess LM. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Med Inform Decis Mak 2021;21(1):54.

DOI: 10.55415/deep-2023-0019.v1

17. Deo RC. Machine Learning in Medicine. Circulation 2015;132(20):1920-1930.
18. Garcia-Garcia E, Gonzalez-Romero GM, Martin-Perez EM, Zapata Cornejo ED, Escobar-Aguilar G, Cardenas Bonnet MF. Real-World Data and Machine Learning to Predict Cardiac Amyloidosis. Int J Environ Res Public Health 2021;18(3).
19. Lv H, Yang X, Wang B, Wang S, Du X, Tan Q, et al. Machine Learning-Driven Models to Predict Prognostic Outcomes in Patients Hospitalized With Heart Failure Using Electronic Health Records: Retrospective Study. J Med Internet Res 2021;23(4):e24996.
20. !!! INVALID CITATION !!! [23, 24].
21. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science 2015;349(6245):255-260.
22. Abukmeil M, Ferrari S, Genovese A, Piuri V, Scotti F. A Survey of Unsupervised Generative Models for Exploratory Data Analysis and Representation Learning. ACM Comput Surv 2021;54(5):Article 99.
23. Driver HE, Kroeber AL. Quantitative Expression of Cultural Relationships: University of California Press; 1932.
24. Sanche R, Lonergan K. Variable reduction for predictive modeling with clustering. Casualty Actuarial Society Forum; 2006: Citeseer.
25. Cantrell CD. Modern Mathematical Methods for Physicists and Engineers. Cambridge: Cambridge University Press; 2000.
26. Craw S. Manhattan Distance. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning and Data Mining. Boston, MA: Springer US; 2017. p. 790-791.
27. Metcalf L, Casey W. Chapter 2 - Metrics, similarity, and sets. In: Metcalf L, Casey W, editors. Cybersecurity and Applied Mathematics. Boston: Syngress; 2016. p. 3-22.
28. Ratcliffe JG. Euclidean Geometry. Foundations of Hyperbolic Manifolds. Cham: Springer International Publishing; 2019. p. 1-33.
29. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. WIREs Data Mining and Knowledge Discovery 2012;2(1):86-97.
30. Sinaga KP, Yang MS. Unsupervised K-Means Clustering Algorithm. IEEE Access 2020;8:80716-80727.
31. Pham DT, Dimov SS, Nguyen CD. Selection of K in K-means clustering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 2005;219(1):103-119.
32. Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. SODA '07; 2007.
33. Bel-Serrat S, Mouratidou T, Santaliestra-Pasias AM, Iacoviello L, Kourides YA, Marild S, et al. Clustering of multiple lifestyle behaviours and its association to cardiovascular risk factors in children: the IDEFICS study. Eur J Clin Nutr 2013;67(8):848-854.
34. Nguyen LH, Holmes S. Ten quick tips for effective dimensionality reduction. PLoS Comput Biol 2019;15(6):e1006907.
35. Köppen M. The curse of dimensionality. 5th online world conference on soft computing in industrial applications (WSC5); 2000.
36. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 1901;2(11):559-572.

DOI: 10.55415/deep-2023-0019.v1

37. Golub GH, Reinsch C. Singular value decomposition and least squares solutions. Numerische Mathematik 1970;14(5):403-420.
38. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 2016;374(2065):20150202.
39. Schölkopf B, Smola A, Müller K-R. Nonlinear component analysis as a kernel eigenvalue problem. Neural computation 1998;10(5):1299-1319.
40. Melit Devassy B, George S, Nussbaum P. Unsupervised Clustering of Hyperspectral Paper Data Using t-SNE. J Imaging 2020;6(5).
41. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290(5500):2319-2323.
42. Kruskal JB. Nonmetric multidimensional scaling: A numerical method. Psychometrika 1964;29(2):115-129.
43. Peterson MD, Liu D, IglayReger HB, Saltarelli WA, Visich PS, Gordon PM. Principal component analysis reveals gender-specific predictors of cardiometabolic risk in 6th graders. Cardiovasc Diabetol 2012;11:146.
44. Biau G, Scornet E. A random forest guided tour. TEST 2016;25(2):197-227.
45. Cortes C, Vapnik V. Support-vector networks. Machine learning 1995;20(3):273-297.
46. Breiman L. Random Forests. Machine Learning 2001;45(1):5-32.
47. Fratello M, Tagliaferri R. Decision trees and random forests. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics 2018;1:3.
48. Winham SJ, Freimuth RR, Biernacka JM. A weighted random forests approach to improve predictive performance. Statistical Analysis and Data Mining: The ASA Data Science Journal 2013;6(6):496-505.
49. Lakshminarayanan B, Roy DM, Teh YW. Mondrian forests: Efficient online random forests. Advances in neural information processing systems 2014;27.
50. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The annals of applied statistics 2008;2(3):841-860.
51. Yan D, Chen A, Jordan MI. Cluster forests. Computational Statistics & Data Analysis 2013;66:178-192.
52. Clémençon S, Depecker M, Vayatis N. Ranking forests. Journal of Machine Learning Research 2013;14:39-73.
53. Zhao Y, Chen F, Zhai R, Lin X, Wang Z, Su L, et al. Correction for population stratification in random forest analysis. Int J Epidemiol 2012;41(6):1798-1806.
54. Aryal S, Alimadadi A, Manandhar I, Joe B, Cheng X. Machine Learning Strategy for Gut Microbiome-Based Diagnostic Screening of Cardiovascular Disease. Hypertension 2020;76(5):1555-1562.
55. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circ Res 2017;121(9):1092-1101.
56. Kawasaki T, Kidoh M, Kido T, Sueta D, Fujimoto S, Kumamaru KK, et al. Evaluation of Significant Coronary Artery Disease Based on CT Fractional Flow Reserve and Plaque Characteristics Using Random Forest Analysis in Machine Learning. Acad Radiol 2020;27(12):1700-1708.
57. Kwak S, Everett RJ, Treibel TA, Yang S, Hwang D, Ko T, et al. Markers of Myocardial Damage Predict

DOI: 10.55415/deep-2023-0019.v1

Mortality in Patients With Aortic Stenosis. J Am Coll Cardiol 2021;78(6):545-558.

58. Yang H, Li X, Cao H, Cui Y, Luo Y, Liu J, et al. Using machine learning methods to predict hepatic encephalopathy in cirrhotic patients with unbalanced data. Comput Methods Programs Biomed 2021;211:106420.
59. Tang J, Henderson A, Gardner P. Exploring AdaBoost and Random Forests machine learning approaches for infrared pathology on unbalanced data sets. Analyst 2021;146(19):5880-5891.
60. Jakkula V. Tutorial on support vector machine (svm). School of EECS, Washington State University 2006;37(2.5):3.
61. Dietrich R, Opper M, Sompolinsky H. Statistical Mechanics of Support Vector Networks. Physical Review Letters 1999;82(14):2975-2978.
62. Szedmak S, Shawe-Taylor J, Saunders CJ, Hardoon DR. Multiclass classification by l1 norm support vector machine. Pattern recognition and machine learning in computer vision workshop; 2004.
63. Xia X-L, Li K. A sparse multi-class least-squares support vector machine. 2008 IEEE International Symposium on Industrial Electronics; 2008: IEEE.
64. Olivier C, Bernhard S, Alexander Z. Transductive Support Vector Machines. Semi-Supervised Learning: MIT Press; 2006. p. 105-117.
65. Datta S, Das S. Near-Bayesian Support Vector Machines for imbalanced data classification with equal or unequal misclassification costs. Neural Netw 2015;70:39-52.
66. Sun W, Chang C, Long Q. Bayesian Non-linear Support Vector Machine for High-Dimensional Data with Incorporation of Graph Information on Features. Proc IEEE Int Conf Big Data 2019;2019:4874-4882.
67. Kim MJ. Building a Cardiovascular Disease Prediction Model for Smartwatch Users Using Machine Learning: Based on the Korea National Health and Nutrition Examination Survey. Biosensors (Basel) 2021;11(7).
68. Petrazzini BO, Chaudhary K, Marquez-Luna C, Forrest IS, Rocheleau G, Cho J, et al. Coronary Risk Estimation Based on Clinical Data in Electronic Health Records. J Am Coll Cardiol 2022;79(12):1155-1166.
69. Krogh A. What are artificial neural networks? Nat Biotechnol 2008;26(2):195-197.
70. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 1943;5(4):115-133.
71. Bebis G, Georgiopoulos M. Feed-forward neural networks. IEEE Potentials 1994;13(4):27-31.
72. Medsker LR, Jain L. Recurrent neural networks. Design and Applications 2001;5:64-67.
73. Gruau F. Automatic definition of modular neural networks. Adaptive behavior 1994;3(2):151-183.
74. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digital signal processing 2018;73:1-15.
75. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. Pattern recognition 2018;77:354-377.
76. O'Shea K, Nash R. An introduction to convolutional neural networks. arXiv preprint arXiv:151108458 2015.
77. Morales MA, van den Boomen M, Nguyen C, Kalpathy-Cramer J, Rosen BR, Stultz CM, et al. DeepStrain: A Deep Learning Workflow for the Automated Characterization of Cardiac Mechanics. Front Cardiovasc Med 2021;8:730316.

DOI: 10.55415/deep-2023-0019.v1

78. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 2015;31(5):761-763.
79. Lih OS, Jahmunah V, San TR, Ciaccio EJ, Yamakawa T, Tanabe M, et al. Comprehensive electrocardiographic diagnosis based on deep learning. Artif Intell Med 2020;103:101789.
80. Narula S, Shameer K, Salem Omar AM, Dudley JT, Sengupta PP. Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography. J Am Coll Cardiol 2016;68(21):2287-2295.
81. Group SR, Wright JT, Jr., Williamson JD, Whelton PK, Snyder JK, Sink KM, et al. A Randomized Trial of Intensive versus Standard Blood-Pressure Control. N Engl J Med 2015;373(22):2103-2116.
82. Crown WH. Real-World Evidence, Causal Inference, and Machine Learning. Value Health 2019;22(5):587-592.
83. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan [website on the Internet]. [cited 2023 Jan 25th]. Available from: https://www.fda.gov/media/145022/download.
84. Raja P, Thangavel K. Missing value imputation using unsupervised machine learning techniques. Soft Computing 2020;24(6):4361-4392.
85. Hasan MK, Alam MA, Roy S, Dutta A, Jawad MT, Das S. Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). Informatics in Medicine Unlocked 2021;27:100799.
86. Quer G, Arnaout R, Henne M, Arnaout R. Machine Learning and the Future of Cardiovascular Care: JACC State-of-the-Art Review. J Am Coll Cardiol 2021;77(3):300-313.
87. Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. Eur Heart J 2019;40(24):1975-1986.
88. Dong X, Yu Z, Cao W, Shi Y, Ma Q. A survey on ensemble learning. Frontiers of Computer Science 2020;14(2):241-258.
89. Chen C-H, Tanaka K, Kotera M, Funatsu K. Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. Journal of Cheminformatics 2020;12(1):19.
90. Deng L, Yu D. Deep Learning: Methods and Applications. Foundations and Trends® in Signal Processing 2014;7(3–4):197-387.
91. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436-444.
92. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform 2017;73:14-29.
93. Spadea MF, Maspero M, Zaffino P, Seco J. Deep learning based synthetic-CT generation in radiotherapy and PET: A review. Med Phys 2021;48(11):6537-6566.
94. Craik A, He Y, Contreras-Vidal JL. Deep learning for electroencephalogram (EEG) classification tasks: a review. J Neural Eng 2019;16(3):031001.
95. Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep Learning for ECG Analysis: Benchmarks and

DOI: 10.55415/deep-2023-0019.v1

Insights from PTB-XL. IEEE J Biomed Health Inform 2021;25(5):1519-1528.

96. Krittanawong C, Johnson KW, Rosenson RS, Wang Z, Aydar M, Baber U, et al. Deep learning for cardiovascular medicine: a practical primer. Eur Heart J 2019;40(25):2058-2073.

DOI: 10.55415/deep-2023-0019.v1

Table 1. Distance measures.

Distance	Measurement *
Euclidean distance	$D(x,y)=\sqrt{\sum_{i=1}^{N}\left(x_i-y_i\right)^2}$
Manhattan Distance	$D(x,y)=\sum_{i=1}^{N}\lvert x_i-y_i\rvert$
Chebyshev Distance	$D(x,y)=\max_i\left(\lvert x_i-y_i\rvert\right)$
Minkowski distances	$D(x,y)=\left(\sum_{i=1}^{N}\lvert x_i-y_i\rvert^p\right)^{\frac{1}{p}}$

* x, y present samples with N variables. x_i, y_i are the numerical value of i^{th} variable. p can be a constant or ∞.

DOI: 10.55415/deep-2023-0019.v1

Table 2. Summary of machine learning algorithm applications in cardiovascular diseases.

ML methods	Investigator	Application
K-means clustering & Hierarchical clustering	Bel-Serrat et al.	Association of multiple lifestyle behaviors and CVD risk
principal component analysis	Peterson et al.	Establishment of cardiometabolic risk patterns with multi
Random forest	Aryal et al.	Prediction of gut microbiome-based diagnostic screening
Support vector machine	Kim et al.	Prediction of prevalence of cardiovascular disease usi [34].
Neural networks	Narula et al.	Discrimination of hypertrophic cardiomyopathy from phy annotated speckle-tracking echocardiographic datasets [3