Machine learning for cardiology

A Review of Machine Learning for Cardiology

Yasir ARFAT[1]*, Gianluca MITTONE[1], Roberto ESPOSITO[1], Barbara CANTALUPO[1], Gaetano M. DE FERRARI[2,3], Marco ALDINUCCI[1]

[1] Computer Science Department, University of Turin, Turin, Italy;
[2] Division of Cardiology, Cardiovascular and Thoracic Department, Città della Salute e della Scienza, Turin, Italy;
[3] Cardiology, Department of Medical Sciences, University of Turin, Turin, Italy.
* Corresponding author: Yasir ARFAT, Computer Science Department, University of Turin, Turin, Italy. e-mail: yasir.arfat@unito.it

Abstract

This paper reviews recent cardiology literature and reports how Artificial Intelligence Tools (specifically, Machine Learning techniques) are being used by physicians in the field. Each technique is introduced with enough details to allow the understanding of how it works and its intent, but without delving into details that do not add immediate benefits and require expertise in the field. We specifically focus on the principal Machine Learning based risk scores used in cardiovascular research. After introducing them and summarizing their assumptions and biases, we discuss their merits and shortcomings. We report on how frequently they are adopted in the field and suggest why this is the case based on our expertise in Machine Learning. We complete the analysis by reviewing how corresponding statistical approaches compare with them. Finally, we discuss the main open issues in applying Machine Learning tools to cardiology tasks, also drafting possible future directions. Despite the growing interest in these tools, we argue that there are many still underutilized techniques: while Neural Networks are slowly being incorporated in cardiovascular research, other important techniques such as Semi-Supervised Learning and Federated Learning are still underutilized. The former would allow practitioners to harness the information contained in large datasets that are only partially labeled, while the latter would foster collaboration between institutions allowing building larger and better models.

Keywords: Cardiology, Machine Learning, Risk Factors, Statistics, Mortality,

1 Introduction

Recent years have witnessed a Cambrian explosion of tools and techniques able to tackle problems that were only solvable by humans up to a few years ago; collectively, we refer to these computer science methods as Artificial Intelligence (AI). AI is accumulating astounding successes at a breakneck pace in both research and applications: from helping in recovering photos by their descriptions[1] on devices used by billions of people to providing tools for investigating the depths of the visible universe[2], AI has never been as capable and popular as today. AI encompasses a vast variety of different techniques: intelligent agents[3], symbolic and subsymbolic reasonings[4], planning[5], case-based reasoning[6], fuzzy systems[7], and expert systems[8] are just a few of them. Despite this diversity, one sub-field in AI single-handedly provided the tools that allowed most of the mentioned successes to be achieved: Machine Learning (ML).

Author's copy (preprint) of Arfat Y, Mittone G, Esposito R, Cantalupo B, De Ferrari GM, Aldinucci M. A review of machine learning for cardiology. Minerva Cardiol Angiol 2021 Aug 02. DOI: 10.23736/S2724-5683.21.05709-4

In this paper, we review some of the recent cardiology literature and report about how ML tools are being used by medical doctors and scientists in the complex tasks of understanding and predicting patients' clinical situations. AI, and specifically ML, can provide clinicians powerful tools supporting and helping everyday crucial clinical decisions[9–11]. For this, the exploitation of AI in medicine is a research direction actively endorsed by national and European funding bodies. The 15M€ EU IA "DeepHealth"[12] (Deep-Learning and HPC to Boost Biomedical Applications for Health, 2019-22) and 6M€ EU RIA "Brainteaser" (BRinging Artificial INTelligencE home for a better cAre of amyotrophic lateral sclerosis and multiple SclERosis, 2021-24) projects are just two recent examples of multi-disciplinary projects directly addressing the development of novel ML tools for AI-assisted diagnosis through medical imaging. With such a great deal of investments and with the renewed interest in the field, there are good chances that AI techniques could become crucial tools to assist clinicians to accurately assess all the relevant factors leading to a diagnosis and the actions that follow. In this context, physicians will remain central to all decisions but supported by tools tailored to ease some of the burdens they face when dealing with the complexity of their work.
ML encompasses all AI approaches specifically oriented towards building models that improve their performances on a given task with experience, that is, data; it is a vast research field able to tackle many tasks and including a vast array of techniques. One broad way to categorize such techniques is by looking at the kind of supervision the learning algorithm receives with the learning examples. The main distinction here is between supervised learning (where all examples are associated with a label), unsupervised learning (where none of the examples is associated with a label), and semi-supervised learning (where only a few of the examples are labeled).
The topic is addressed from a technical perspective, introducing criteria to compare the different techniques, explaining them, and critically reviewing their pros and cons. We describe and review the most important risk scores based on ML techniques, allowing the reader to have a comprehensive perspective on the AI applications currently available in the cardiovascular (CV) field. We also briefly analyze the main statistical approaches, comparing them with ML methods.
Figure 1 gives an overview of the paper structure, describing in particular Sections 3 and 4, summarising the usual process followed by ML practitioners: the data are first preprocessed to improve their quality (removing missing values, performing feature selection, …), then the ML algorithm is trained on them. The model obtained as output is iteratively refined by searching for optimal (hyper) parameters by performing experiments on the training or the validation data. Finally, the model is evaluated on the test data and deployed for usage. After that Section 5 deals with statistical methods, Section 6 discusses our findings and Section 7 concludes the paper.

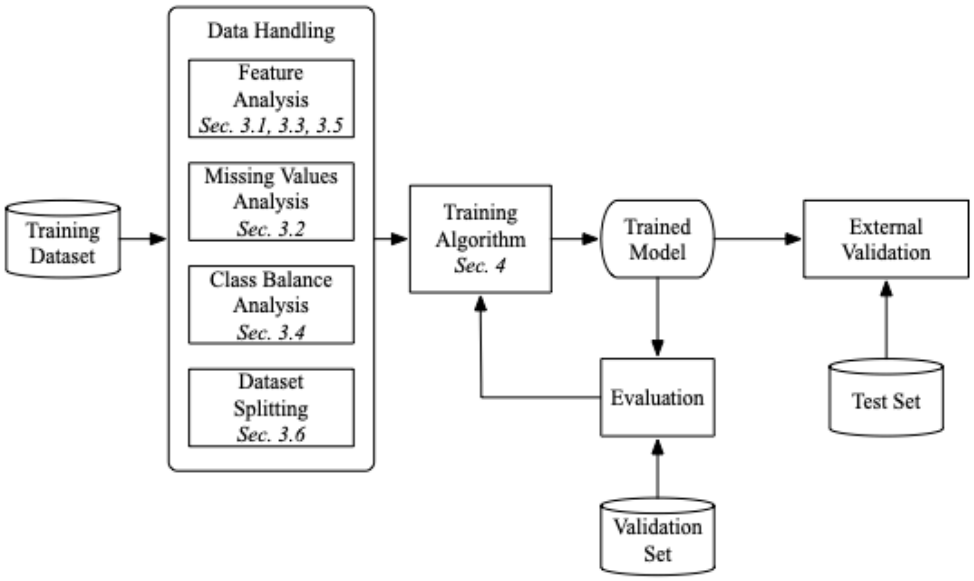

Figure 1 Figure 1. The machine learning process.

2 Methodology

This work is a state-of-the-art review, meaning that its primary goals are to address the current knowledge in the field of AI-based cardiological risk scores and offer new perspectives on their development. The latest papers were collected reporting reviews and comparisons of current methodologies from Google Scholar, covering a wide range of works in computer science and cardiology communities. From there, we proceeded backward, exploring the main literature strands focusing specifically on supervised learning. This process led to reading 58 recent papers; the most influential ones (based on the quality of obtained results, practical applications' usefulness, and the sophistication of the techniques exploited) have been selected, summarized, and described (see Section 4).

As the reader shall see, only supervised techniques are reviewed. While we initially set out to include unsupervised and semi-supervised methods in our review, we realized that researchers in the CV field are not currently exploiting these techniques. We comment on this aspect in the final discussion (see Section 6).

3 Data handling

In the supervised learning scenario data comes from a labeled dataset $X = \{(x_i, y_i)\}$ where examples (a.k.a. samples) x_i are associated with labels y_i and are assumed to be i.i.d. (independent and identically distributed). This section discusses widespread data preprocessing techniques to overcome common issues like outliers, missing values, noisy readings, and many others that often affect the learning performance.

3.1 Features types

In the medical scenario, the samples x_i usually describe patients' data and are structured into several fields known as features in the ML community. Features can take many forms, but as far as most learning algorithms are concerned, they can be subdivided into three different categories[13]:

- quantitative: those with a meaningful numerical scale;
- ordinal: ordered features without a scale;
- categorical: those without an ordering or scale.

The feature type is crucial to ML algorithms since not all algorithms can deal with all kinds of features, and even when they can, they usually handle them differently. Some feature types are more informative than others: quantitative features contain more details than the ordinal ones, and the same relationship holds between ordinal and categorical features. The empirical impact of this statement is present in many of the papers included in this review: risk scores obtained by reducing the number of used features often end up using more quantitative and ordinal features than categorical ones[14–24].

3.2 Missing values

An aspect that is important to discuss further is the handling of missing values. Many different approaches exist to deal with this problem, relying on and exploiting different assumptions on the meaning of a missing value. In the medical field, the absence of a value can have a significant clinical meaning; if some values are not collected, there could be some specific reason[25] (e.g., the medical treatment prevents data from being collected, or some values are derived from others). In those cases, expert intervention is needed to understand how to handle the issue correctly. For some models, like Decision Trees, a correct approach to address this issue can be creating a specific value for missing data, signifying that data could not be collected, giving more information to the model than the simple missing value.

If data is not missing for a specific reason, imputation can be exploited to guess its value; this is a powerful technique capable of enhancing the richness of information in a dataset, but it should be carefully handled since it can drastically reduce the data variance. Imputation is frequently exploited

in the field[14–16,21,26–28], especially employing Monte Carlo or regression methods. Some works explicitly targeting the imputation of medical data are also available[29,30].

3.3 Feature selection

While it is intuitive that the more features are available, the more precise the prediction will be, this is not always the case. On the one hand, by adding more features to the training process, the ones related to the target will more likely be available to the ML algorithm; but, on the other hand, the risk of capturing random regularities grows exponentially with the number of added features. A high number of features makes the predictive process also less interpretable. In addition, from a medical perspective, it is not useful to introduce multiple features referring to the same medical parameter: these will be highly correlated and will not add any relevant information to the process. Feature selection is a way to overcome these problems and can be achieved in various ways.
The most frequently used approach for feature selection that we found in the reviewed literature is the forward selection/backward elimination process (as, for instance in papers:[14–17,31–34]), in which the model is trained multiple times using different sets of features. At each iteration, features are added/eliminated according to a greedy strategy. Other strategies are available[35] each of them addressing specific situations.

3.4 Class imbalances

A common issue in medical datasets is the balance between the investigated classes of patients[36]. Since ML models try to optimize some kind of misclassification loss, when the classes are very imbalanced the algorithm may decide that it is better to simply disregard (or to focus less of its efforts on) the minority class since errors on that class do not contribute too much to the classification error anyway. This is a problem, and it should be taken into account when working with imbalanced datasets (this also holds for other ML tasks like regression and clustering). In this scenario, it is appropriate to counter the problem to ensure that the algorithm reaches its full potential in terms of generalization capabilities[17,20,24,37]. There are two standard techniques for achieving this: oversampling (duplicate samples from minority class) or undersampling (removal of samples from the majority class); while the first approach can lead to some bias if data duplication is not correctly applied, the second approach inevitably leads to loss of information.

3.5 Feature normalization

One more technique that can be exploited to obtain better performance with some models like K-Nearest Neighbors, Support Vector Machines (SVMs), Naive Bayes, and Neural Networks (see Section 4 for an introduction to these models) is feature normalization. It consists of rescaling all the numerical features to have them on the same scale, thus allowing the ML algorithms that exploit numerical methods (e.g., gradient descent, distance-based algorithm) to work better way[18,19,27,38–40].
We can apply feature normalization in several ways. In cases where the feature values are all positive, one can scale them to the [0,1] range by dividing each value by their maximum; otherwise, it is common to scale so that the values have zero mean and unit variance by subtracting the mean and dividing the result by the standard deviation of the feature being normalized.

3.6 Dataset splitting

After data have been pre-processed to enhance their quality, they should be prepared for the learning process. Typically, the whole dataset is divided into two or three different smaller sets. The ML names for these sets are:

- training set: data used to train the models;
- validation set: data used to tune the hyperparameters of the model;
- test set: data used to assess the generalization capability of the model.

In medical literature, these terms are sometimes different:

- derivation cohort/set corresponds to the training set;
- no specific cohort/set are identified explicitly for hyperparameters tuning;

- validation cohort/set corresponds to the test set.

Following the three splits schema allows to correctly train, tune and evaluate the ML model without compromising the rigorousness of the results. Many of the reviewed paper authors do not use a validation set[21,27,28,40–46] tuning the hyperparameters of their models on the training set or the test set. It is worth emphasizing that using only two splits is not considered a best practice, because one is likely to overfit either the training set or (worse) the test set.

3.7 Dataset size

The majority of the datasets used in the current studies span from few thousands[14–19,47–49] to hundreds of thousands of patients[20,27,41,46], while it is unusual to find smaller ones[24,33,40,43]. The general trend is to use ever larger and larger datasets over time: this is positive, since the dataset size requirements grow with the "complexity" of the concepts to be learned. Conducting an ML study only on a few hundred patients would severely limit the scope of possible applications. In this regard, it is worth mentioning that, in particular cases, useful knowledge can be extracted from a limited amount of data by exploiting a deep knowledge on the specific phenomena to be analyzed[19,20,23,44]. The data sources can be either a local trial[31] or a shared resource like a national or international registry[14–16,32].

3.8 Follow-up time

A fixed follow-up time for the investigations makes the learning process more effective, resulting in less data variance and more interpretable results. It is also possible to incorporate time in the predictive process, but this type of analysis is more complex and delicate. Some works exploit this technique to obtain survival time predictions or time-to-event analysis; for this kind of analysis, statistical techniques are typically more effective than ML ones. For instance, the SHFM risk score[32] is based on the Cox Proportional Hazard Model, explored in section 5, and takes time into account for his inference. The most well-known counterpart in ML is DeepHit[50], a Deep Neural Network-based survival analysis tool; its first application to medical data appears now in some preprints.

3.9 Privacy, security, and features

Typically, medical datasets include a list of clinical features like age, sex, type of diabetes, etc.; it is not unusual to include patients' habits like smoking or drinking. These are all sensitive information that shall be manipulated according to privacy policies: many techniques allow handling sensitive data without the need to share it or move it physically (e.g., edge computing[51] and federated learning[52]). Still, we are currently unaware of any study where such technologies are exploited in the CV field.

4 Machine Learning techniques

In this section, the most common supervised techniques are introduced. A summary of them is shown in Figure 2, and references to relevant literature are provided in Table 1. Table 2 provides a list of references organized according to these study objectives. We shall explain the differences between different techniques and contextualize their usage in the current literature. To do that, we need a few tools to make high-level but grounded claims about the techniques themselves.

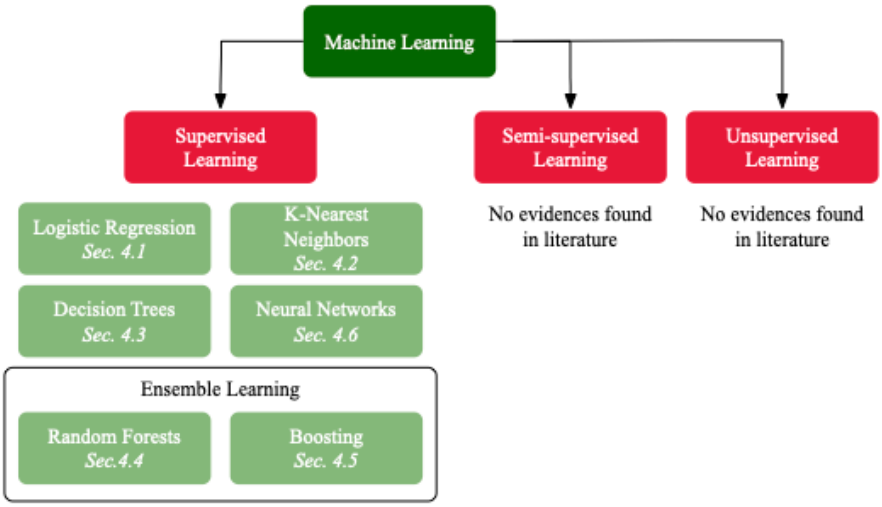

Figure 2 Machine Learning techniques discussed in this paper.

Specifically, we will need to discuss two essential ML concepts: the "bias/variance decomposition of the error" and the "learning bias" of learning algorithms. Since the term "bias" is used with slightly different meanings in these two topics, we shall use the "bias" or "bias component of the error" to denote the former sense (the one used in the bias/variance decomposition) and always use the term "learning bias" for the latter.

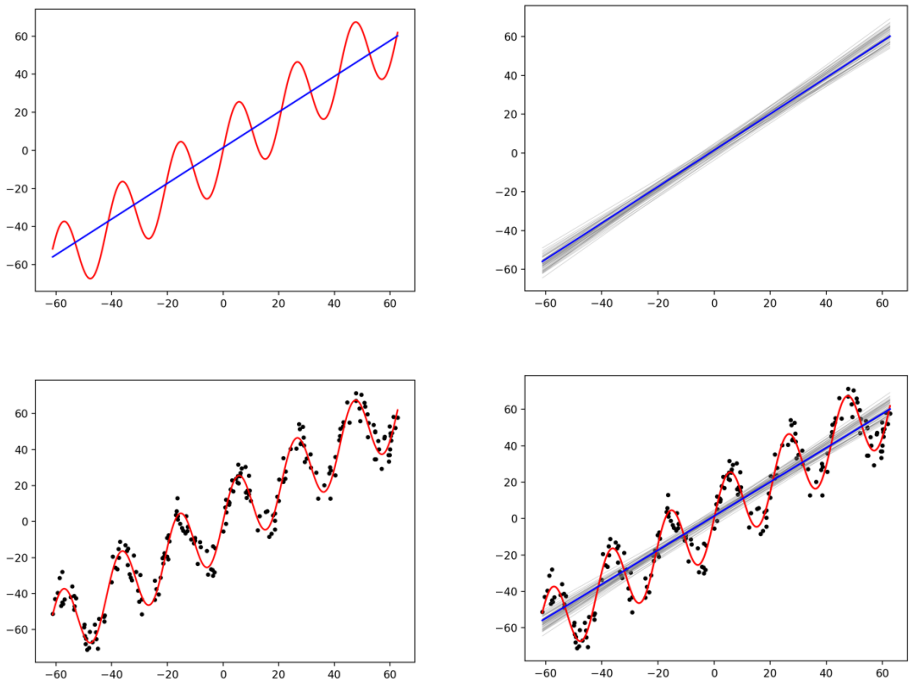

Figure 3 Examples of (a) bias, (b) variance, (c) noise decomposition, and (d) aggregated bias, variance and noise. The red curved line is the true concept to be approximated, the blue line is the average regressor, gray lines are individual regressors, and black dots are noisy observations. As can be seen, these three error components have a massive effect on approximation performance.

Let us start from the bias/variance decomposition of the error[53,54]. In a nutshell: the error made, on average, by a learning algorithm can always be decomposed into three components: bias, variance, and noise (see Figure 3). The bias component of the error measures how much the average decision surface differs from the true concept. This difference usually correlates with the concept space size

searched by the algorithm: if the algorithm searches between all linear concepts and the true concept is a cubic polynomial, then the bias component of the error will be significant. The variance part of the error measures how much, on average, a concept learned by the algorithm differs from the average concept. The variance component of the error usually correlates inversely with the size of the concept space searched by the algorithm: in our previous example, a learning algorithm exploring the space of all cubic polynomial will have a larger error variance than that of an algorithm searching a linear concept space (having more degree of freedom, it will more easily adapt to variations in the dataset, thus producing more diversified results). Noise is just the error component due to errors in acquiring the example's features or labels.
Bias and variance are usually competing forces, and decreasing one often causes the other to increase. Knowing if an algorithm has high/low bias/variance allow one to understand which are the best possible actions to improve results and enable to compare algorithms based on how brittle they are (how much the variance component of their error is high) and how well they would work when combined with other methods (e.g., ensemble techniques).
The learning bias of an algorithm refers to the heuristic that the learning algorithm adopts to choose between different concepts. This broad definition encompasses many details of the algorithm, such as the space it searches and how it selects between equally good concepts on the training set. Understanding the algorithm's learning bias is important because it is the key to understand how much it is suited for the problem at hand. Indeed the no-free-lunch theorem[55] implies that, without further assumptions, all learning algorithms are created equal and perform equally (bad/well) on a random problem. In other terms, there is no better/best algorithm in absolute terms: an algorithm is only as good as the fitness of its learning bias on the problem at hand.

4.1 Logistic Regression

Logistic Regression (LR) is a supervised algorithm that can induce models for classification tasks[1]. The main assumption made by the algorithm (i.e., its learning bias) is that the logarithm of the odds $log\left(\frac{P(y=1)}{P(y=0)}\right)$is a linear function of the input x. The implication, which gives the name to the technique, is that the probability of the positive class has the form of the logistic function $f(x) = \frac{e^{x}}{1+e^{x}}$ (see Figure 4). Formally:

$$log\left(\frac{P(y=1)}{P(y=0)}\right) = wx$$
$$\Rightarrow log\left(\frac{P(y=1)}{1-P(y=1)}\right) = wx$$
$$\Rightarrow \frac{P(y=1)}{1-P(y=1)} = e^{wx}$$
$$\Rightarrow P(y=1) = e^{wx} - P(y=1)e^{wx}$$
$$\Rightarrow P(y=1) = \frac{e^{wx}}{1+e^{wx}}$$

The main benefits of logistic regression are that, being a linear model, it tends to have low variance and requires small sample sizes to perform well. For the same reasons, it does not usually work well when the relationships to be modeled are not linear (in that case, the higher bias component of the error tends to be not compensated enough by the low variance).

[1]Please note that the "regression" part in the name of the technique can be misleading. The name is due to the fact that the main idea is to predict the probability of the classes (which is a numeric value which justifies the regression name), but it is then almost always used to solve classification problems.

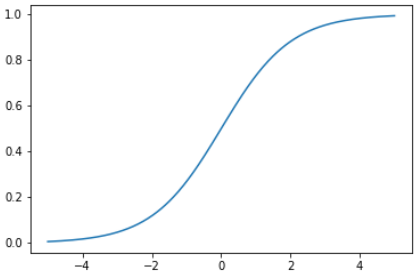

Figure 4 The logistic function.

Among the technologies we found in the papers we reviewed, LR ranks very high in popularity and performance. HAS-BLED[56] aims to provide a simple score for major bleeding risk in patients with atrial fibrillation. In this case, LR has been combined with univariate statistical analysis to iteratively select a subset of features highly correlated with the risk of major bleeding. EuroSCORE II[57] was built exploiting univariate LR, likelihood ratios, and Akaike's Information Criterion to select a subset of features highly correlated with the investigated endpoint (cardiac surgical mortality). These features are then inserted in a logistic equation that gives the final predicted mortality. This study is an update of the original 1999 EuroSCORE[58] and exploits a broad international database of patients. More advanced techniques are used in Iung et al.[59], where LR is exploited in univariate and multivariate fashion and in a hierarchical way to study data heterogeneity across different medical centers. In SPRM[60] a multinomial LR is used on a selected set of features (obtained by univariate LR, variance, χ^2 analysis, and backward elimination) to directly compare the proportion of mortality attributable to two distinct causes: information that other models can hardly provide. ScREEN[42] exploits LR to do feature selection and the Youden index for establishing cut-offs for quantitative variables. Each selected feature was assigned a single risk point, and the total risk of collateral events after a Cardiac Resynchronization Therapy is the sum of these points. Many other papers exploring this topic exist[27,45], but they simply apply the above LR techniques to other pathologies to the best of our knowledge.

4.2 K-Nearest Neighbors

Despite its age and simplicity, K-Nearest Neighbors (KNN) is still used in some of the works we reviewed. The main idea of the KNN algorithm is to store the dataset in memory and then compute the predictions for a new example x by recovering the K examples nearest to x and averaging the results (for regression) or deciding the class by a majority vote (for classification). Here the learning bias is in the assumption that similar examples (as estimated by the distance measure adopted) should have similar labels. From the bias/variance decomposition point of view, this algorithm has some flexibility since the K parameter controls the trade-off (the lower is the K parameter, the lower is the bias component, and the higher the variance component). In Elsayed et al.[40] KNN has been used to detect the early risk of coronary artery diseases.

4.3 Decision Trees

Decision trees[61] (DTs) are tried and tested ML tools that can be easily applied to a wide variety of problems and provide very interpretable results. DTs' flexibility derives from their capability to do both classification and regression and their ability to work well with a wide variety of feature kinds (numerical, categorical, ordinal, …). As mentioned, DTs are very interpretable models since they can be interpreted as a list of nested if-then-else clauses. They are very brittle models, meaning that they are low bias, high variance models. They also tend to overfit data unless countermeasures are taken; for these reasons, DTs are usually pruned (making the trees smaller, lowering their variance

at the expense of a higher bias) and averaged using some form of ensemble algorithm. Pruning introduces an additional learning bias, which is to prefer shorter trees over taller ones.
While DTs are still very popular when used as pieces of an ensemble model (e.g., as components of Random Forests (4.4) or used as weak learners in Boosting procedures (4.5)), they are not very popular as standalone models. In Furui et al.[43] long-time prediction for atrial fibrillation has been performed using DTs: the researchers produced a risk stratification system using a classification and regression tree, categorizing the patients in low, medium, and high risk.

4.4 Random Forest

Random Forest (RF) is an ensemble learning supervised machine learning algorithm. Its flexibility, performance, and ease of use make it a very popular choice in many application contexts for regression and classification tasks. Ensemble learning encompasses different techniques where many models are combined to build a more robust one; in RF, this technique is Bagging[62]. This algorithm creates many copies of a model, each one of them being trained on a different subset of the available data and combines them through a simple majority vote or averaging the predictions. Bagging's main strength is the ability to reduce the variance of the combined models, i.e., it works best with low-bias, high-variance algorithms[63,64]. RF builds on these strengths by bagging models obtained by a slightly updated version of the DTs learning algorithm that exacerbates these traits of DT models.
Our literature review found that RF appeared second highest in the papers we reviewed, ranging from classification tasks[26,33,46] to regression tasks. Many authors use RFs to predict all-cause mortality, and others apply RF to particular cardiovascular diseases[37,46] or for risk assessment of heart failure[33] and venous thromboembolism[22]. In some studies, authors claim that the inferred RF models are helpful for clinical decisions, allowing to estimate whether a patient is suffering from heart failure with preserved ejection fraction or not[33], and that their risk score assessment performs better than the state-of-the-art ones.[22,26,46]

4.5 Boosting

Boosting algorithms are particular kinds of ensemble algorithms. Boosting algorithms encompass those ensemble techniques that guarantee a decrease in the training error (usually by descending the gradient of some loss suffered by the ensemble). These models are very popular since they are typically easy to use, have very few parameters, can be applied to both classification and regression tasks, and tend not to overfit the data. Several boosting algorithms have proved to be particularly popular in our literature review: AdaBoost, LogitBoost, Gradient Boosting, Light Gradient Boosting, and eXtreme Gradient Boosting.
Adaboost[65] is the original boosting algorithm, and it can be shown to optimize the exponential loss suffered by the ensemble implicitly. LogitBoost[66] is a variant of AdaBoost derived by casting AdaBoost as a generalized additive model and substituting the cost function with the logistic loss $\sum_i \quad log(1 + e^{-y_i f(x_i)})$ (where i sums over all examples (x_i, y_i)). Gradient Boosting is a variant of boosting where the loss function is explicitly optimized via gradient descent, and eXtreme Gradient Boosting is a refined version of the Gradient Boosting approach[67].
Recently, AdaBoost has been used to predict all-cause mortality[18,47] and report accuracies on-par or better than the state-of-the-art based on LR[42] and provide clues useful for the clinical decision-making process. LogitBoost has been instrumental in developing cardiovascular risk predictions[17,48], outperforming established risk scores such as the Framingham Risk Score and the Segment Stenosis Score.
eXtreme Gradient Boosting has been applied in predicting mortality[38,41] and in predicting the risk of cardiovascular disease Coronary Artery Calcium Score[49]. Ye et al.[41] introduced a risk assessment tool based on this latter technique that provides early prediction of older people's mortality using Electronic Health Record. In van Rosendael et al.,[68] authors also used eXtreme Gradient Boosting to enhance the risk stratification to maximize coronary CTA usage derived from plaque information. In all these works, the authors reported that using eXtreme Gradient Boosting their overall accuracy improved with respect to the competing approaches.

Li[34] has applied Light Gradient Boosting to Intensive Care Unit patients on data collected from three hospitals. The authors claim that their approach helps make better clinical decisions, and the model achieves good performance for predictions.
Gradient Boosting of DTs has been used with success in several interesting works trying to predict mortality[24] and heart failures[39]. The latter work also provides additional information for medical staff to understand complications after heart failure.

4.6 Neural Networks

Neural Networks (NNs) are connectionist models with roots in cybernetics and the attempt to model the human brain[69]; since then, the models evolved into practical tools with only a faint resemblance to the first ones developed. After an exciting start in the ’60s and a resurgence in the ’80s, NNs did not progress for almost two decades. For many tasks, they required too much data and computational power, and the research focused on simpler models that were easier to train. Thanks to breakthrough in the training algorithms, progresses in CPUs and GPUs, the creation of big computational farms[70], and the advent of internet tools allowing the collection of huge labeled datasets, NNs have seen new interest from the research community and are nowadays at the forefront of the research in many critical applicative fields.
NNs are built from basic units called neurons, which can be easily arranged in layers. Layers are easy to connect, and the whole network can be trained end-to-end using the Stochastic Gradient Descent algorithm[71]. While a single-layer network can approximate any function to arbitrary precision (as implied by the Universal Approximation Theorem[72]), the real power of these models is in the automatic abstractions provided by stacking multiple layers into a Deep Neural Network (DNN). Each layer abstracts its inputs providing the following layer with a data representation that is easier to work within the context of the task being solved. State-of-the-art NNs models are DNNs models and have been shown to provide super-human performances[73,74] on many tasks involving hard-to-abstract data such as those involved with image and audio processing.
Shallow NNs have been used to predict mortality due to heart failure[20,23] showing performances outperforming other learning methods despite being trained on unbalanced datasets in recent literature. DNNs have been exploited for predicting the risk of mortality[19,75] or heart failure and acute heart failure[44]. The authors of these two works compared DNNs with other ML techniques showing performance improvements.
Another interesting recent application of NNs in this field is to exploit their ability to work with data correlated very complicatedly. This is the case of the Deep Cox Mixtures[76], in which a NN assists a Cox Regression Model (Section 5.1) to fit the hazard ratios of the regression. This work is based on a sound statistical and ML background, comprehensively exposed, and offers state-of-the-art performance when working with different groups of individuals.

5 Statistical approaches

Alongside the ML approaches discussed above, it is worth briefly describing the main statistical techniques currently used in the prediction of cardiovascular events since they still cover an essential role in the field[77,78]. As in the previous section, every major technique implied in the field will be reviewed and explained, together with a brief comment on the top risk scores that exploit them.

5.1 The Cox Proportional Hazards Model

Survival Analysis is a broad branch of statistics that studies how much time it takes for an event to occur, or, in the specific case of medical applications, how much time would likely pass before an event affects a given individual. Given this brief description of survival analysis, it is clear that its tools are well-fit for predicting medical events. In this scenario, the Cox Proportional Hazards (PH) Model[79] takes a predominant place as being, by far, the most used statistical technique for the prediction of cardiovascular events[14,15,21,28,31,32,80].

The typical analysis of the relation between a single risk factor and an event is carried on by evaluating the Instantaneous Hazard Rate $\lambda(t)$. This measure is defined as the rate at which events occur, given the total number of individuals at risk. Formally:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{Ev(t, t+\Delta t)/N(t)}{\Delta t}$$

where $Ev(t, t+\Delta t)$ is the number of events occurring between times t and $t+\Delta t$ and $N(t)$ is the number of individuals at risk at time t.
Typically, in medical studies, authors are interested in comparing different populations, each with different characteristics, like the assumption (or the lack thereof) of a drug. It is then common to model each group's survival possibilities to assess the effects of the given drug on the population. In these cases, it is convenient to model the variations in risk hazards of the different populations by mean of the Hazard Ratio (HR), which is the ratio of the two different Instantaneous Hazard Rates:

$$HR = \frac{\lambda_1(t)}{\lambda_0(t)}$$

where $\lambda_0(t)$ and $\lambda_1(t)$ are the Instantaneous Hazard Rate for the populations 0 and 1 at time t. With an HR ratio above 1, the events are more likely to occur in population one and vice-versa, with the magnitude of HR indicating the difference's strength.
The kind of analysis shown before, although helpful, can be applied only to investigate the impact of a single risk factor on the survival possibilities of a population. More complex tools are needed to assess the simultaneous impact of multiple risk factors. In this context, the Cox PH model finds its bases, allowing the assessment of the effects of multiple risk factors on the survival time of a population under three assumptions[81]:

- the survival capability of an individual is independent of the other individuals in the population;
- the risk factors and the hazard are multiplicatively related (i.e., incrementing one of the risks multiplies the hazard);
- the HR over time is constant.

These assumptions make the Cox PH model semi-parametric since it assumes the relationship between risk factors and hazard but not on the hazard function itself. A conditional argument by Cox justifies this approach, but it is not presented here. In general, a Cox PH model can be written as:

$$\lambda(t|X_i) = \lambda_0(t) exp(X_i \beta)$$

where X_i is the vector of the risk factor values for the ith individual (usually called covariates in this context), β is the vector of regression coefficients, and $\lambda_0(t)$ is the baseline hazard when all the risk factors are zero. The values assessed for every risk factor (βs) impact on the population's survival; positive values of β will proportionally increase the hazard risk and vice-versa. Of course, it is possible to calculate the HR of two hazard rates calculated with a Cox PH model; it is then possible to investigate the survival capabilities of different populations based on multiple risk factors.

Heart Failure Survival Score

One of the most well-known risk scores exploiting a Cox PH model is the Heart Failure Survival Score (HFSS)[31]. The first step for the derivation of this score has been the clinical features selection by mean of univariate statistical analysis methods like Kaplan-Meier method[82] and log-rank tests[83]; in this way, the researchers successfully reduced the analysis on a set of forty features, against the eighty available. The Cox PH model has been applied to these features, but with two additional strategies: a stepwise forward-entry/backward-elimination selection, based on the p-value, and best-subset discovery, based on a χ^2 test. In this way, a subset of only eleven features has been selected as the best trade-off between feature number and predictive power. The HFSS is then defined as the absolute value of the sum of the products of the Cox PH model coefficients and the respective risk factor value ($|\beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_n x_n|$where $x_0, x_1, \ldots, x_n$ are the actual variable values and $\beta_0, \beta_1, \ldots, \beta_n$ are the computed coefficients). This risk score achieved good results for its time, but it lacked generalization capabilities: performance was limited when applied to other datasets than

the one on which it was derived, due to the low number of patients involved in the study and the specific requirements that they had to match. A positive aspect of this work, though, is that not only one model has been developed: two of them have been derived, one exploiting an invasive medical feature (mean PCWP) and the other one not; despite that the two models reached similar performance, thus raising the attention on the opportunity of (not) performing invasive procedures.

Seattle Heart Failure Model

Another risk score exploiting the Cox PH model is the Seattle Heart Failure Model (SHFM)[32]. In this case, the feature selection has been made by means only of the Cox PH model, with a stepwise forward-entry/backward-elimination, partially from the derivation dataset, partially from large published trials (for the features not exhaustively described by the derivation cohort). Once the model has been derived, the SHFM score is then defined as the sum of the products of the β-coefficients with the value of the corresponding parameter ($SHFM\ score = |\beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_n x_n|$). The survival value at time t for a patient is then defined as $survival(t) = e^{(-\lambda t) e^{(SHFM\ score)}}$, where $e^{-(\lambda t)}$ is the baseline survival (survival at time t when all risk factors are zero) and λ the slope/year derived from the dataset. Due to how it is constructed, this risk score allows a per-patient analysis, and his reliability is well-documented since it has been tested on five different datasets; it is also an example of a score in which some risk factors (like age and sex) have been forced into the model, thus merging the statistical approach and the medical knowledge.

ORBIT

In ORBIT[14] the Cox PH model is used, together with a feature selection step based on the backward selection process, to create the best performing model based on a pool of medically relevant risk factors for major bleeding. The derivation dataset is large, counting more than ten thousand patients, and the missing data were imputed only once through Markov Chain Monte Carlo or regression methods. From the full final model, only five risk factors have been retained, the ones with the highest χ^2 statistic, and to each one of them, an integer score is assigned, based on the strength of their correlation with major bleeding. The result is a simple risk score, easily computable in a real-world situation. This risk score is another example of how a limited group of risk factors can be exploited to obtain good performance. To assess the technique's performance, the paper reports an evaluation on an external dataset where ORBIT is compared against HAS-BLED[84] and ATRIA[85].
The GISSI[80] risk score exploits a similar approach, but in this case, there are not predefined integer points assigned to each final feature, but a nomogram is provided for bedside application; in this way, it is possible also to take into account the value of the risk predictors.

PARIS

PARIS[15] is another risk score based on the Cox PH model: differently from the other approaches presented above, it exploits data imputation in the derivation process. Employing a multivariate normal regression, specific missing values of decisive risk factors have been imputed multiple times and, for each set of imputed data, a Cox PH model with backward selection has been fitted to the data. These different models are used to obtain a fully calibrated final Cox PH model. From that model, the β-coefficient is used to obtain integer values for the risk factors. This approach has been repeated two times, one for the derivation of the major bleeding model and the other for the coronary thrombotic event one, allowing physicians to evaluate the risk of these two events through two integer risk scores.

PRECISE-DAPT

The PRECISE-DAPT[16] score exploits the Cox PH model in two flavors, both univariate and multivariate, with backward elimination, to assess the potential predictors of major and minor TIMI bleeding. The result is an integer risk score computed on five clinical variables. Each variable is associated with an integer score based on its value and its β-coefficient. The paper also offers a nomogram for bedside single-patient evaluation. This score has been derived on a broad dataset and

validated on two external cohorts. It has also been compared to the PARIS score during the evaluation to assess the difference between these two approaches.

6 Discussion

Figure 5 reports the histogram of the frequencies with which ML techniques have been used in the papers we reviewed. While the figure shows quite a number of different approaches, we observe that most of the experimentation happens with ensemble techniques. In fact, 9 techniques (RF, AdaBoost, Gradient Boosting, eXtreme Gradient Boosting, Light Gradient Boosting, LogitBoost, Gradient Boosting of DTs, Explainable Boosting Machines, and CatBoost) out of 16 are ensemble algorithms. It also happens that while Random Forests is the most popular ensemble approach, most of the others are some variants of Gradient Boosting. The preference for ensemble learning in general, and Boosting, in particular, is highly understandable since Boosting usually gets very accurate models without requiring much tuning of the parameters. Also, Boosting naturally counterbalances overfitting[86] by increasing the "margin" of the classification even when the training error stops decreasing.

If we remove ensemble learning from the picture, we see that Logistic Regression is used almost as much as all the remaining approaches cumulatively (14 times versus 17). Again, this technique is straightforward to apply and very robust to overfitting so that it can be easily applied to smaller datasets. The remaining models (in order of popularity) are SVMs, Naive Bayes, and KNN. SVMs, despite being very popular in our sample of papers, never achieve the highest ranking, often being outperformed by Logistic Regression. In theory, SVMs should be able to match or outperform logistic regression when properly configured and trained. Unfortunately, there is often no information22,44 or very little information19,22,33,39 in the papers we reviewed about how SVMs are configured (which kernels are used in the experiments and how other hyperparameters have been chosen), so it is hard to tell how much effort was devoted to tuning these tools to the problem at hand.

Naive Bayes and KNN are seldom used, and they do not seem to perform well anyway; KNN ranks first in one case[40], but in that case, the study has only 60 patients and compares it only with one other approach (Random Forest).

Figure 6 reports, for each technique, the number of times it ranked first in the papers where it competed. In that figure, we omit to count an algorithm as ranking first if the paper did not compare it with other methods (explaining why some of the methods appear with a count of 0 and the sum of the counting is shorter than the list of papers we reviewed). Despite not being a very popular method, in the three occasions where NNs are used, they perform best. This is not a surprising result: NNs are notoriously hard to train and bring the necessity of selecting many hyperparameters, which explains why they are not the preferred choice in many works. However, when they are properly configured and when data is abundant, they usually perform very well.

As mentioned at the beginning of this paper, we focused our work on supervised learning techniques. This was a compulsory choice since we could not find any recent paper mentioning unsupervised or semi-supervised learning (SSL) techniques. Given the cost and difficulty of labeling large datasets, we believe that SSL techniques[87] could be a very useful research direction. In semi-supervised learning, only a small percentage of the examples are given a label, i.e., one can think to have access to two datasets, one small labeled dataset $D_l = \{(x_i, y_i)\}$ and one large unlabeled dataset. The algorithms then exploit the information in D_l and the structure of the distribution $P(x)$ inferred from D_u either to induce labels for D_u itself without trying to infer a general labelling rule (this is called the transductive setting); or to induce a general rule for labeling future examples (inductive setting). It comes without saying that whenever this works well, it introduces huge benefits in terms of accuracy of the models and in costs of tools development. However, the applicability of these techniques necessitates that the $P(x)$ distribution has some good structure. Examples of nice properties of the $P(x)$ distribution are:

1. the examples are high-dimensional, but they live on a low-dimensional manifold;
2. examples are clustered and belonging to the same cluster implies having similar labels;
3. nearby examples are likely to have similar labels.

Whether or not any of these assumptions apply entirely depends on the domain at hand and a good research direction could be to study if and in what measure they apply on cardiology prediction tasks.

Working with medical data is difficult. Not only for the magnitude of problem complexities but also because, due to privacy concerns, it is not easy for researchers to share the data they collect. This causes a fragmentation of the available data that prevents training very complex models such as DNNs. In this regard, we believe that the recent advancements in privacy-preserving techniques such as (to name just a few) Federated Learning[88] and Differential Privacy[89] could provide help. In a nutshell: Federated Learning allows one to train models locally and to combine them globally without ever revealing the local datasets. Differential Privacy gives formal guarantees that the contents of a dataset cannot be inferred by exploiting the inferred model. The two methods combined allow one to train global models without moving data while also guaranteeing against privacy attacks on the inferred global model.

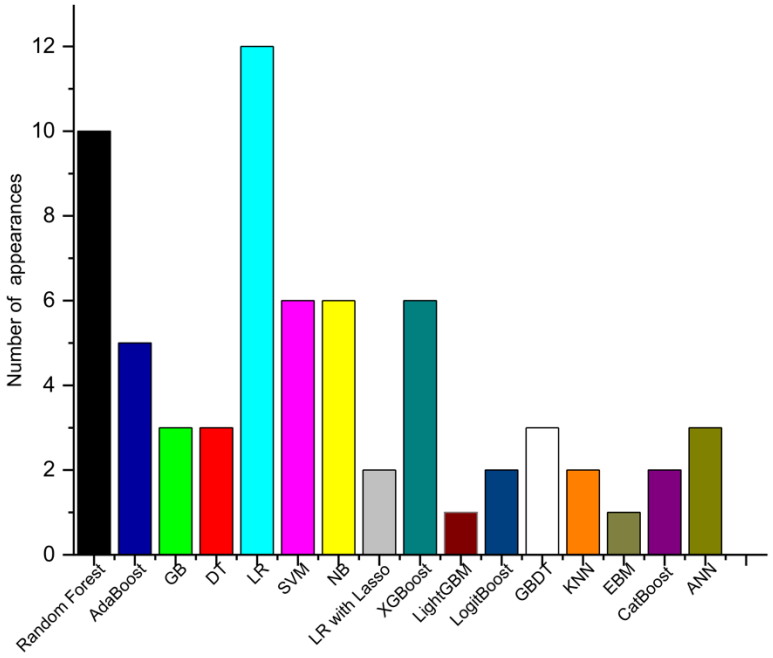

Figure 5 Number of appearances of Machine Learning techniques in the reviewed literature. Abbreviations: Random Forest (RF),Gradient Boosting (GB), Decision Trees (DT), Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), eXtreme G

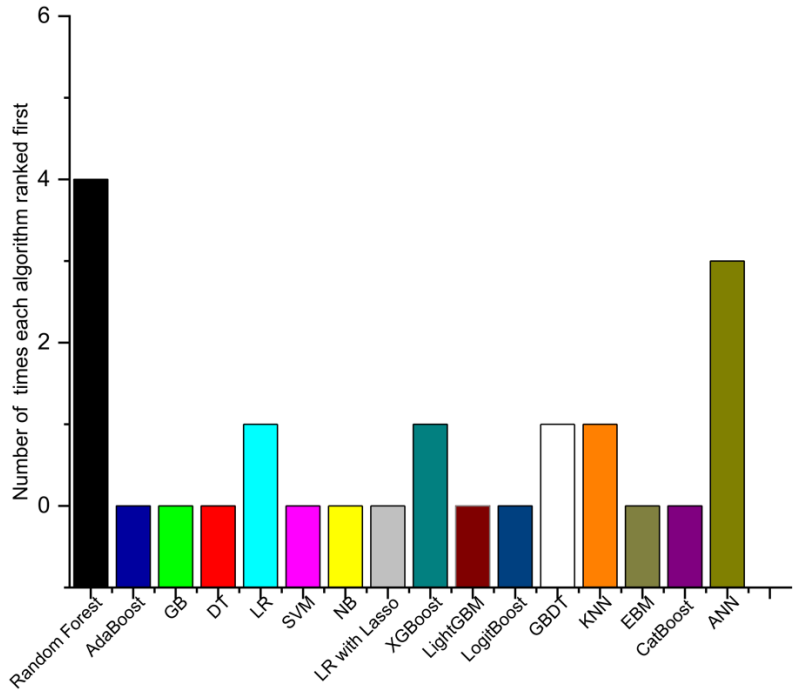

Figure 6 Number of times each Machine Learning technique ranked first in the reviewed literature. Papers where only a single technique was presented are not included. Abbreviations: Random Forest (RF),Gradient Boosting (GB), Decision Trees (DT), Logist

7 Conclusions

AI is a growing force in everyday life, and its usage is slowly but consistently percolating in medical professions. In most cases, ML is the main force driving the adoption of AI. In this paper, we presented a review of the main applications of ML in recent cardiology literature. We provided an introduction of the techniques used most often, reviewed competing statistical techniques, and critically reviewed these tools' usage in recent applications and research. We found that in most cases, the usage of ML is limited to tools that have been firmly understood for many years. In our opinion, newer and more data-hungry approaches are currently under-represented. Indeed, one of the problems we outline is the paucity of very large datasets. In fact, the efforts to build such datasets are hampered by difficulties in labeling vast amounts of data and privacy concerns that do not allow merging datasets acquired by different institutions. We suggested exploiting semi-supervised techniques to tackle the labeling problem and combine Federated Learning and Differential Privacy to overcome privacy issues.

Medical data is challenging to collect, usually noisy, and the involved tasks are hard to solve. ML can be very useful to ease the burden on physicians, and, in part, it is already helping in that area. We hope that, with improvements in data collection and their sharing, better models will be learned; physicians will be able to work faster and more accurately, and, ultimately, many lives will be saved.

REFERENCES

1. Pham H, Dai Z, Xie Q, Luong M-T, Le QV. Meta Pseudo Labels. *ArXiv200310580 Cs Stat*. Published online March 1, 2021. Accessed June 22, 2021. http://arxiv.org/abs/2003.10580

2. Kremer J, Stensbo-Smidt K, Gieseke F, Pedersen KS, Igel C. Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy. *IEEE Intell Syst*. 2017;32(2):16-22. doi:10.1109/MIS.2017.40

3. Brenner W, Zarnekow R, Wittig H. *Intelligent Software Agents: Foundations and Applications*. Springer Science & Business Media; 2012.
4. Garcez A, Besold TR, Raedt L, et al. Neural-symbolic learning and reasoning: contributions and challenges. Published online 2015.
5. Wilkins DE. Practical Planning: Extending the Classical AI Planning Paradigm. Elsevier; 2014.
6. López B. Case-based reasoning: a concise introduction. *Synth Lect Artif Intell Mach Learn*. 2013;7(1):1-103.
7. Yager RR, Zadeh LA. *An Introduction to Fuzzy Logic Applications in Intelligent Systems*. Vol 165. Springer Science & Business Media; 2012.
8. Liebowitz J. The Handbook of Applied Expert Systems. cRc Press; 2019.
9. Dorado-Díaz PI, Sampedro-Gómez J, Vicente-Palacios V, Sánchez PL. Applications of artificial intelligence in cardiology. The future is already here. *Rev Esp Cardiol Engl Ed*. 2019;72(12):1065-1075.
10. Bonderman D. Artificial intelligence in cardiology. *Wien Klin Wochenschr*. 2017;129(23):866-868.
11. Sardar P, Abbott JD, Kundu A, Aronow HD, Granada JF, Giri J. Impact of artificial intelligence on interventional cardiology: from decision-making aid to advanced interventional procedure assistance. *Cardiovasc Interv*. 2019;12(14):1293-1303.
12. Colonnelli I, Cantalupo B, Merelli I, Aldinucci M. StreamFlow: cross-breeding cloud with HPC. *IEEE Trans Emerg Top Comput*. Published online 2020.
13. Flach P. Machine Learning: The Art and Science of Algorithms That Make Sense of Data. Cambridge University Press; 2012.
14. O'Brien EC, Simon DN, Thomas LE, et al. The ORBIT bleeding score: a simple bedside score to assess bleeding risk in atrial fibrillation. *Eur Heart J*. 2015;36(46):3258-3264.
15. Baber U, Mehran R, Giustino G, et al. Coronary thrombosis and major bleeding after PCI with drug-eluting stents: risk scores from PARIS. *J Am Coll Cardiol*. 2016;67(19):2224-2234.
16. Costa F, van Klaveren D, James S, et al. Derivation and validation of the predicting bleeding complications in patients undergoing stent implantation and subsequent dual antiplatelet therapy (PRECISE-DAPT) score: a pooled analysis of individual-patient datasets from clinical trials. *The Lancet*. 2017;389(10073):1025-1034.
17. Han D, Kolli KK, Gransar H, et al. Machine learning based risk prediction model for asymptomatic individuals who underwent coronary artery calcium score: Comparison with traditional risk prediction approaches. *J Cardiovasc Comput Tomogr*. 2020;14(2):168-176.
18. D'Ascenzo F, De Filippo O, Gallone G, et al. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. *The Lancet*. 2021;397(10270):199-207.
19. Guimarães P, Keller A, Böhm M, et al. Risk prediction with office and ambulatory blood pressure using artificial intelligence. *medRxiv*. Published online 2020.
20. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Dwivedi G. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. *ESC Heart Fail*. 2019;6(2):428-435.
21. Pocock SJ, Ariti CA, McMurray JJ, et al. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *Eur Heart J*. 2013;34(19):1404-1413.
22. Wang X, Yang Y-Q, Liu S-H, Hong X-Y, Sun X-F, Shi J. Comparing different venous thromboembolism risk assessment machine learning models in Chinese patients. *J Eval Clin Pract*. 2020;26(1):26-34.
23. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Chow BJ, Dwivedi G. Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. *PloS One*. 2019;14(6):e0218760.
24. Mamprin M, Zinger S, de With PHN, Zelis JM, Tonino PAL. Gradient boosting on decision trees for mortality prediction in transcatheter aortic valve implantation. In: *Proceedings of the 2020 10th International Conference on Biomedical Engineering and Technology*. ; 2020:325-329.
25. Schmidt D, Niemann M, von Trzebiatowski GL. The Handling of Missing Values in Medical Domains with Respect to Pattern Mining Algorithms. In: *CS&P*. ; 2015:147-154.
26. Tokodi M, Schwertner WR, Kovács A, et al. Machine learning-based mortality prediction of patients undergoing cardiac resynchronization therapy: the SEMMELWEIS-CRT score. *Eur Heart J*. 2020;41(18):1747-1756.
27. O'Sullivan JW, Shcherbina A, Justesen JM, et al. Combining clinical and polygenic risk improves stroke prediction among individuals with atrial fibrillation. *medRxiv*. Published online 2020.
28. Vazquez R, Bayes-Genis A, Cygankiewicz I, et al. The MUSIC Risk score: a simple method for predicting mortality in ambulatory patients with chronic heart failure. *Eur Heart J*. 2009;30(9):1088-1096.
29. Cismondi F, Fialho AS, Vieira SM, Reti SR, Sousa JM, Finkelstein SN. Missing data in medical databases: Impute, delete or classify? *Artif Intell Med*. 2013;58(1):63-72.

30. Janssen KJ, Donders ART, Harrell Jr FE, et al. Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol*. 2010;63(7):721-727.
31. Aaronson KD, Schwartz JS, Chen T-M, Wong K-L, Goin JE, Mancini DM. Development and prospective validation of a clinical index to predict survival in ambulatory patients referred for cardiac transplant evaluation. *Circulation*. 1997;95(12):2660-2667.
32. Levy WC, Mozaffarian D, Linker DT, et al. The Seattle heart failure model. *Circulation*. 2006;113(11):1424-1433.
33. Angraal S, Mortazavi BJ, Gupta A, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC Heart Fail*. 2020;8(1):12-21.
34. Li X, Xu X, Xie F, et al. A Time-Phased Machine Learning Model for Real-Time Prediction of Sepsis in Critical Care. *Crit Care Med*. 2020;48(10):e884-e888.
35. Jović A, Brkić K, Bogunović N. A review of feature selection methods with applications. In: *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. Ieee; 2015:1200-1205.
36. Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. *Int J Mach Learn Comput*. 2013;3(2):224.
37. Ricciardi C, Edmunds KJ, Recenti M, et al. Assessing cardiovascular risks from a mid-thigh CT image: a tree-based machine learning approach using radiodensitometric distributions. *Sci Rep*. 2020;10(1):1-13.
38. Kilic A, Goyal A, Miller JK, et al. Predictive utility of a machine learning algorithm in estimating mortality risk in cardiac surgery. *Ann Thorac Surg*. 2020;109(6):1811-1819.
39. Chen Y, Qi B. Representation learning in intraoperative vital signs for heart failure risk prediction. *BMC Med Inform Decis Mak*. 2019;19(1):1-15.
40. Elsayed HAG, Syed L. An automatic early risk classification of hard coronary heart diseases using framingham scoring model. In: *Proceedings of the Second International Conference on Internet of Things, Data and Cloud Computing*. ; 2017:1-8.
41. Ye C, Li J, Hao S, et al. Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. *Int J Med Inf*. 2020;137:104105.
42. Providencia R, Marijon E, Barra S, et al. Usefulness of a clinical risk score to predict the response to cardiac resynchronization therapy. *Int J Cardiol*. 2018;260:82-87.
43. Furui K, Morishima I, Morita Y, et al. Predicting long-term freedom from atrial fibrillation after catheter ablation by a machine learning algorithm: Validation of the CAAP-AF score. *J Arrhythmia*. 2020;36(2):297-303.
44. Kwon J, Kim K-H, Jeon K-H, et al. Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. *PloS One*. 2019;14(7):e0219302.
45. Hernandez-Suarez DF, Kim Y, Villablanca P, et al. Machine learning prediction models for in-hospital mortality after transcatheter aortic valve replacement. *Cardiovasc Interv*. 2019;12(14):1328-1338.
46. Yang L, Wu H, Jin X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. *Sci Rep*. 2020;10(1):1-8.
47. Adler ED, Voors AA, Klein L, et al. Improving risk prediction in heart failure using machine learning. *Eur J Heart Fail*. 2020;22(1):139-147.
48. Motwani M, Dey D, Berman DS, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J*. 2017;38(7):500-507.
49. Lee J, Lim J-S, Chu Y, et al. Prediction of Coronary Artery Calcium Score Using Machine Learning in a Healthy Population. *J Pers Med*. 2020;10(3):96.
50. Lee C, Zame W, Yoon J, van der Schaar M. Deephit: A deep learning approach to survival analysis with competing risks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol 32. ; 2018.
51. Satyanarayanan M. The emergence of edge computing. *Computer*. 2017;50(1):30-39.
52. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. *ACM Trans Intell Syst Technol TIST*. 2019;10(2):1-19.
53. Domingos P. A unified bias-variance decomposition. In: *Proceedings of 17th International Conference on Machine Learning*. ; 2000:231-238.
54. Kohavi R, Wolpert DH. Bias plus variance decomposition for zero-one loss functions. In: *ICML*. Vol 96. ; 1996:275-283.
55. Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Comput*. 1997;1(1):67-82.
56. Pisters R, Lane DA, Nieuwlaat R, De Vos CB, Crijns HJ, Lip GY. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. *Chest*. 2010;138(5):1093-1100.
57. Nashef SA, Roques F, Sharples LD, et al. Euroscore ii. *Eur J Cardiothorac Surg*. 2012;41(4):734-745.

58. Nashef SA, Roques F, Michel P, et al. European system for cardiac operative risk evaluation (Euro SCORE). *Eur J Cardiothorac Surg*. 1999;16(1):9-13.
59. Iung B, Laouénan C, Himbert D, et al. Predictive factors of early mortality after transcatheter aortic valve implantation: individual risk assessment using a simple score. *Heart*. 2014;100(13):1016-1023.
60. Shadman R, Poole JE, Dardas TF, et al. A novel method to predict the proportional risk of sudden cardiac death in heart failure: derivation of the Seattle Proportional Risk Model. *Heart Rhythm*. 2015;12(10):2069-2077.
61. Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1(1):81-106.
62. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123-140.
63. Bühlmann P, Yu B. Analyzing bagging. *Ann Stat*. 2002;30(4):927-961.
64. Esposito R, Saitta L. Monte Carlo theory as an explanation of bagging and boosting. In: *IJCAI*. Vol 3. ; 2003:499-504.
65. Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: *Icml*. Vol 96. Citeseer; 1996:148-156.
66. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann Stat*. 2000;28(2):337-407.
67. Mason L, Baxter J, Bartlett P, Frean M. Boosting algorithms as gradient descent in function space. In: Nips; 1999.
68. van Rosendael AR, Maliakal G, Kolli KK, et al. Maximization of the usage of coronary CTA derived plaque information using a machine learning based algorithm to improve risk stratification; insights from the CONFIRM registry. *J Cardiovasc Comput Tomogr*. 2018;12(3):204-209.
69. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev*. 1958;65(6):386.
70. Aldinucci M, Rabellino S, Pironti M, et al. HPC4AI: an AI-on-demand federated platform endeavour. In: *Proceedings of the 15th ACM International Conference on Computing Frontiers*. ; 2018:279-286.
71. Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat*. Published online 1951:400-407.
72. Baker MR, Patil RB. Universal approximation theorem for interval neural networks. *Reliab Comput*. 1998;4(3):235-239.
73. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *nature*. 2016;529(7587):484-489.
74. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE International Conference on Computer Vision*. ; 2015:1026-1034.
75. Jallepalli S, Pathak P, Gupta P, Kar S, Gupta M. Development and Validation of Artificial Intelligence-Based Cardiovascular Disease (AI-CVD) Risk Score. *Available SSRN 3444410*. Published online 2019.
76. Nagpal C, Yadlowsky S, Rostamzadeh N, Heller K. Deep Cox mixtures for survival regression. *ArXiv Prepr ArXiv210106536*. Published online 2021.
77. Canepa M, Fonseca C, Chioncel O, et al. Performance of prognostic risk scores in chronic heart failure patients enrolled in the European Society of Cardiology Heart Failure Long-Term Registry. *JACC Heart Fail*. 2018;6(6):452-462.
78. Yoshida R, Ishii H, Morishima I, et al. Performance of HAS-BLED, ORBIT, PRECISE-DAPT, and PARIS risk score for predicting long-term bleeding events in patients taking an oral anticoagulant undergoing percutaneous coronary intervention. *J Cardiol*. 2019;73(6):479-487.
79. Cox DR. Regression Models and Life Tables," Journal of the Royal Statistical Society. Scr. B, 34. 187-202.(1975). *Partial Likelihood Biom*. 1972;62:269-276.
80. Barlera S, Tavazzi L, Franzosi MG, et al. Predictors of mortality in 6975 patients with chronic heart failure in the Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico-Heart Failure trial: proposal for a nomogram. *Circ Heart Fail*. 2013;6(1):31-39.
81. Harrell FE. Cox proportional hazards regression model. In: *Regression Modeling Strategies*. Springer; 2015:475-519.
82. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457-481.
83. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *J R Stat Soc Ser Gen*. 1972;135(2):185-198.
84. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*. 2010;137(2):263-272.
85. Fang MC, Go AS, Chang Y, et al. A new risk scheme to predict warfarin-associated hemorrhage: The ATRIA (Anticoagulation and Risk Factors in Atrial Fibrillation) Study. *J Am Coll Cardiol*. 2011;58(4):395-401.

86. Bartlett P, Freund Y, Lee WS, Schapire RE. Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann Stat*. 1998;26(5):1651-1686.
87. Van Engelen JE, Hoos HH. A survey on semi-supervised learning. *Mach Learn*. 2020;109(2):373-440.
88. Kairouz P, McMahan HB, Avent B, et al. Advances and open problems in federated learning. *ArXiv Prepr ArXiv191204977*. Published online 2019.
89. Dwork C, Roth A. The algorithmic foundations of differential privacy. *Found Trends Theor Comput Sci*. 2014;9(3-4):211-407.

Appendix: A guide to references by topic

Table 1: References by topic

Row	References	Topic
1	14,15,16,17,18,19,20,21,22,23,24	Risk scores that reduce the number of used features often end up using more quantitative and ordinal features than categorical ones.
2	14,15,16,21,26,27,28	Imputation is a technique that is frequently exploited in the field.
3	14,15,16,17,19,21,23,28,31,32,33,34	Forward selection/backward elimination process for feature selection.
4	17,18,20,24,28,37	Measures to counter the class imbalance problem.
5	18,19,27,38,39,40	Features normalization techniques
6	21,27,28,40,41,42,43,44,45,46	Splitting criteria in CV papers
7	14,15,16,17,18,19,23,32,34,37,42,44,47,48,49	Studies with thousands of patients
8	20,27,41,46	Studies with hundreds of thousands of patients
9	24,26,31,33,42,43,40	Studies with smaller datasets

Table 2: Summary of Machine Learning and Statistical techniques found in literature

References	Technique	Short Summary
27, 42, 45, 56,57, 58, 59, 60,	Logistic Regression	Very popular ML technique which implements a linear with sound theoretical underpinning. *Pros*: stable, low variance, model. Works well even with small datasets. *Cons*: it is a linear model, will not work well when more complex relationships need to be captured.
40	K-Nearest Neighbors (KNN)	Very simple model with some applications in cardiology literature. A new example is labeled according to the labels of its nearest neighbors in the training data. *Pros*: can be used for both regression and classification tasks; easy to interpret; training time is negligible (it only needs to save the dataset somewhere); the K parameter allows for tuning the bias/variance tradeoff. *Cons*: performance problems at test time due to high usage of memory, and expensive computation when the dataset is large; no theoretically sound way to set K, needs to be estimated by cross-validation.
43	Decision Trees	Very popular machine learning technique, very widely available in free and commercial ML tools. The learnt model is a tree having tests on attributes in the internal nodes and label decisions in the leaves. A decision can be then explained by looking at the result of each test that leads to the leaf used for the decision. *Pros*: very flexible tool; it can be used for classification and regression; explainable decisions. *Cons*: must be regularized to avoid overfitting; not always state-of-the-art accuracy-wise.
22, 26, 33, 37, 46	Random Forest	Ensemble method that is very widely used in cardiology research. It builds a set of decision trees and averages their answers. *Pros*: the averaging of the answers tends to counteract the natural tendency of decision trees to overfit the data. Pros: it usually sports very good performances with minimal efforts. *Cons*: the natural interpretability of decision trees is hampered by the combination of multiple models.
17, 18, 24, 34, 38, 39, 41, 42, 47-49, 68	Boosting	Ensemble method that guarantees, under mild assumptions, to drive training error to zero. Several different techniques goes under this name (e.g., AdaBoost, LogitBoost, Light gradient boosting, and Extreme gradient boosting). *Pros*: usually easy to use; very few parameters; can be applied to both classification and regression tasks; resilient to overfitting. *Cons*: combining multiple models make the decision hard to interpret.
19, 20, 23, 44,75,76	Neural Networks	State-of-the-art models for several tasks that deal with raw signals (e.g., images and videos). They are not widely used in cardiovascular literature. *Pros*: unparalleled ability to deal with images, sounds and videos; they can learn highly non-linear decision surfaces. *Cons*: hard to train; many hyperparameters; on tabular data other techniques are usually easier to apply; work best with very large datasets.
14-16, 21, 28, 31, 32, 79, 80, 81,82,83	Statistical techniques	Statistical techniques for the prediction of cardiological events.

Table 3: Summary of objectives/goals in different research papers

Research Papers	Major Objectives
17,19,20,22-24,26,33,38,41,44,45,47,48,68	Predict all-cause mortality (ACM)
20,23,33,39,42,47	Predict heart failure (HF)
34,43	Improving clinical decision
37,49	Classify different disease
27	Predict risk of stroke
40,48	Predict coronary artery disease