Proposed Requirements for Cardiovascular Imaging Related Machine Learning Evaluation (PRIME) Checklist

Partho Sengupta, Sirish Shrestha, Béatrice Berthon, Emmanuel Messas, Erwan Donal, Geoffrey Tison, James Min, Jan d'Hooge, Jens-Uwe Voigt, Joel Dudley, et al.

Proposed Requirements for Cardiovascular Imaging Related Machine Learning Evaluation (PRIME) Checklist

Reviewed by the American College of Cardiology Health Care Innovation Council

Running Title: The PRIME Checklist

Partho P Sengupta, MD, DM, FACC, FASE[1], Sirish Shrestha, MSc [1], Béatrice Berthon, PhD [2], Emmanuel Messas, MD [3], Erwan Donal, MD [4], Geoffrey Tison, MD, PhD [5], James K Min, MD [6], Jan D'hooge, MD [7], Jens-Uwe Voigt, MD [7], Joel Dudley, PhD [8,9], Johan Verjans, MD, PhD [10, 11], Khader Shameer, PhD [8,9], Kipp Johnson, PhD [8,9], Lasse Løvstakken, PhD [12], Mahdi Tabassian, PhD [7], Marco Piccirilli, PhD [1], Mathieu Pernot, PhD [2], Naveena Yanamala, MS, PhD [1], Nicolas Duchateau, PhD [13], Nobuyuki Kagiyama, MD, PhD [1], Olivier Bernard, PhD [13], Piotr Slomka, PhD[14], Rahul Deo, MD, PhD [5], Rima Arnaout, MD [5]

Address for correspondence:
Partho P. Sengupta, MD, DM
Heart & Vascular Institute,
West Virginia University,
1 Medical Center Drive,
Morgantown, WV 26506-8059.
E-mail: Partho.Sengupta@wvumedicine.org

* The views expressed in this paper do not reflect the views of the American College of Cardiology or the Journal of American College of Cardiology family.

[1]West Virginia University Heart & Vascular Institute, Division of Cardiology, Morgantown, WV, USA; [2]Physique pour la Médecine Paris, Inserm U1273, CNRS FRE 2031, ESPCI Paris, PSL Research University, Paris, France; [3]Université Paris Descartes, Sorbonne Paris Cité, Paris, France; [4]Département de Cardiologie et Maladies Vasculaires, Service de Cardiologie et maladies vasculaires, CHU Rennes, Rennes, France; [5]Division of Cardiology, Department of Medicine, University of California San Francisco, San Francisco, CA, USA; [6]Cleerly, Inc. New York, New York; [7]Laboratory on Cardiovascular Imaging & Dynamics, Department of Cardiovascular Science, KU Leuven, Leuven, Belgium; [8]Department of Genetics & Genomic Sciences and Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [9]Institute for Next Generation Healthcare, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [10]South Australian Health and Medical Research Institute, Adelaide, SA, Australia; [11]Australian Institute for Machine Learning, University of Adelaide, North Terrace, Adelaide, SA, Australia; [12]Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway; [13]CREATIS, CNRS UMR 5220, INSERM U1206, Université Lyon 1, INSA-LYON, France; [14]Department of Imaging and Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA

Disclosures:

Mr. Shrestha has nothing to disclose
Dr. Arnaout has nothing to disclose
Dr. Bernard has nothing to disclose
Dr. Berthon has nothing to disclose
Dr. D'hooge has nothing to disclose
Dr. Deo has nothing to disclose
Dr. Donal has nothing to disclose
Dr. Duchateau has nothing to disclose
Dr. Dudley is employed and has salary support from Tempus Labs.
Dr. Johnson has an industry relationship with Tempus Labs in the form of Salary. He also has an equity interest in Tempus Labs and in Oova, Inc.
Dr. Kagiyama has nothing to disclose
Dr. Lovstakken is a consultant for GE Vingmed Ultrasound AS.
Dr. Messas has nothing to disclose
Dr. Min has employment and salary suport from Clearly Inc. and advisory relationship with Arienta.
Dr. Pernot has nothing to disclose
Dr. Piccirilli has nothing to disclose
Dr. Khader has nothing to disclose
Dr. Slomka has received a grant from Siemens Medical Systems. He has also received royalties from Cedars-Sinai.
Dr. Tabassian has nothing to disclose
Dr. Tison has nothing to disclose
Dr. Verjans has nothing to disclose
Dr. Voigt has nothing to disclose
Dr. Yanamala is member of Think2Innovate LLC.
Dr. Sengupta is a consultant for the following companies: 1) Heartsciences. 2) Ultromics 3) Kencor Health. He also hold equity interests in Ultromics, Kencor Health and Heartsciences.

Abstract

Machine learning (ML) has been increasingly used within cardiology, particularly in the domain of cardiovascular imaging. Due to the inherent complexity and flexibility of ML algorithms, inconsistencies in the model performance and interpretation may occur. Several review articles have been recently published that introduce the fundamental principles and clinical application of ML for general cardiologists. The current document builds on these introductory principles and outlines a more comprehensive list of crucial responsibilities that need to be completed when developing ML models. The document thus aims to serve as a scientific foundation to aid investigators, data scientists, authors, editors, and reviewers involved in machine learning research with the intent of uniform reporting of ML investigations. An independent multidisciplinary panel of ML experts, clinicians, and statisticians worked together to review the theoretical rationale underlying seven sets of requirements that may reduce algorithmic errors and biases. Finally, the document summarizes a list of reporting items as an itemized checklist that highlight steps for ensuring correct application of ML models and the consistent reporting of model specifications and results. It is expected that the rapid pace of research and development and the increased availability of real-world evidence may require periodic updates to the checklist.

Keywords: Machine Learning, Artificial Intelligence, Reporting Items, Checklist, Cardiovascular imaging

Abbreviations:
AI, Artificial intelligence
ML, Machine learning

Highlights:

- Algorithmic complexity and flexibility of ML techniques can result into inconsistencies in model reporting and interpretations.
- The PRIME Checklist provides seven items to be reported for reducing algorithmic errors and biases.
- The checklist aims to standardize reporting on model design, data, selection, assessment, evaluation, replicability and limitations.
- As artificial intelligence and ML technologies continue to grow, the checklist would need periodic updates in future.

In 2016, during a two day Think Tank meeting, The American College of Cardiology's Executive Committee and Cardiovascular Imaging Section Leadership Council initiated a discussion regarding the future of cardiovascular imaging among thought leaders in the field (1). One of the goals was focused on machine learning (ML) tools and methods that embrace data-driven approaches for scientific inquiry. The 2016 document stressed the creation and adoption of standards, the development of registries, and the use of new techniques in bioinformatics. Furthermore, the imaging community's unfamiliarity with the approach was cited as a potential barrier to widespread adoption. In the recent years, the field of cardiac imaging has seen a remarkable burst of innovation with the use of ML, demonstrating powerful algorithms that start impacting the ways clinical and translational research is designed and executed (2-15).

ML is a subfield of artificial intelligence (AI) where an algorithm automatically discovers patterns of data in the datasets without using explicit instructions. Several recent state-of-the-art review articles have focused on providing introductory concepts regarding ML algorithm applications for general cardiologists (3, 8, 16, 17). While ML is creating headlines in medical journals, congress, and on the web, considerable uncertainty and debate have arisen around topics such as problems with real-world data sources, the inconsistent availability of labeled data and outcome information, bias injection, inaccurate measurements, reproducibility, lack of external validation, and insufficient reporting, which contribute to hindering the reliable assessment of prediction model studies and reliable interpretations of the results by clinicians. This Proposed Recommendations for Cardiovascular Imaging Related Machine Learning Evaluation (PRIME) Checklist aims to provide general framework as reference in guiding scientific work for investigators, data scientists, authors, editors, and reviewers involved in machine learning research in cardiovascular imaging. The goal of the PRIME Checklist document is to standardize the application of artificial intelligence (AI) and ML, including data preparation, model selection, and performance assessment. The document provides a set of strategic steps towards developing a pragmatic checklist (Table 1), which may allow consistent reporting of machine learning models in cardiovascular imaging studies. To further determine its ease of use and applicability, we illustrate the application of checklist developed in PRIME Checklist for two recent articles that developed ML models in cardiovascular imaging (Supplementary Table S1).

1. Designing the Study Plan

Defining the goal of the analysis is a key first step that informs many downstream decisions as to whether to use machine learning at all including the presence or absence of labeled data for supervised or unsupervised learning. These informed decisions can alter the approach to model training, model selection, development, and tuning.

Determining the appropriateness of machine learning to the dataset

The first question researchers should address is whether the ML approach could be applicable and beneficial to their study. There is overlap between traditional statistics and ML, but they differ regarding the extent of the assumptions and the formulation of the methods to either predict or make inferences. If the dataset is relatively small (i.e., fewer than hundreds of sample per class for "average" modeling problems), then overfitting becomes a much bigger concern resulting in models that fail to generalize well for unseen instances (18). Similarly, if variables that are important to modeling the data are missing or if the model is too simple, the resulting ML model may underfit the data, thus producing less than optimal results (18). Statistical analyses rely on simpler models that are not necessarily optimized to the specific data under observation, and are therefore less prone to overfitting, at the price of lower performance if the problem is complex (19). On the contrary, datasets with large numbers of features, on the order of thousands or those having many irrelevant or redundant features, with regard to a given task may just as easily lend to overfitting. ML is especially useful where the data are unstructured, feature selection or exploratory analyses are preferred to identify meaningful insights. As such, the learning algorithm may find patterns in the data to generate a homogenous fraction and identify relationships in a data-driven manner beyond the a priori knowledge or existing hypotheses (3). While the use of advanced machine learning algorithms may be better suited for handling big or heterogeneous datasets, it comes at the cost of the interpretability, complexity, and the ability to draw a causal inference. Caution should be taken against causally interpreting results derived from models designed primarily for prediction. For tasks where the goal is to establish causality, the techniques that are commonly used in "traditional" biostatistics, including statistical analyses methods such as propensity score matching, or bayesian inference, maybe better suited; however, newer methods involving ML algorithms are being developed for causal inferences (20, 21).

Understanding and Describing the Data

Irrespective of whether ML tools or statistical analysis methods are used, it is crucial to understand and describe the data available for analysis to draw appropriate conclusions, whether it is tabular, images, time-series data, or a combination. Important considerations about the data include the availability of data that is representative of the target population, the method used to obtain data, and the resultant biases that may influence the conclusions that can be drawn from the data. Describing the data can also help understand the relevance to the target population. The method of data collection, including the sampling method, is also important, as bias may be introduced from systematic error, coverage error, or selection. Various guidelines and associated checklists for medical research have been established to aid in the reporting of relevant details about the data, depending on the study design (22). Clearly describing the data preprocessing or data cleaning methods used is essential to enable reproducibility.

It should be acknowledged that all ML or statistical algorithms are guided by basic data assumptions; an independent and identical distribution is an important assumption where the random variables are mutually independent and have the same statistical distribution and properties. Methods to check for the model assumptions, such as learning curve (23), diagnosing bias and variance (24) or error analyses, may be required.

Defining the process

When building ML models, it is crucial to specify the inputs (e.g., pixels in images, a set of parameter values, and patient information), and desired outputs (e.g., object categories and the presence or absence of disease, an integer representing each category, the probability for each category, the prediction of a continuous outcome measurement, transformed pixel data) that are required. While defining outputs is essential for supervised learning approaches, unsupervised learning approaches may also benefit from defining the output that is desired for the task to select an appropriate model. Some tasks, once well-defined, can only be achieved using certain types of algorithms. For example, image recognition tasks from raw pixel/image data may require the extraction of the optimal features from the data, which is intrinsically performed by deep neural networks and goes beyond the use of hand-crafted features as input. At the conceptual level, deep learning works by breaking a complex task (e.g., identifying a tumor or other abnormalities in organs/tissue) into simple fewer abstract tasks. For example, if the task is to identify a square from other geometric shapes in an image, this task can be divided further into smaller nested sub-steps i.e., by first checking if there are four lines associated with a shape or not. Alternately, one could check if lines are inter-connected and perpendicular to each other and whether they are closed or not and so on in a step-by-step hierarchical fashion. After the consecutive hierarchical identification of complex task, deep-learning approaches automatically find out the features that are important for solving the problem. It is generally preferable to start by defining the overall broader analysis, necessary to accomplish the task of interest, and dividing it into several sub-tasks. Defining the problem or task as precisely as possible can also help guide the data annotation strategy and model selection. Once the data analysis objective has been identified and the inputs/outputs have been defined for each task, it is easier to determine the appropriate models for the analysis pipeline (Figure 1).

Strategic steps for developing the checklist:

- Identify and assess if machine learning could be appropriate.
- Define the objectives of machine learning to achieve the overall goal.
- Understand and describe the data.
- Identify input and target variables.
- Describe the baseline data and understand biases that may exist

2. Data Standardization, Feature Engineering and Learning

Data preparation, standardization and feature extraction is key to the success of model development. It ensures that the data format is appropriate for machine learning, the utilized variables carry relevant information for solving the problem at hand and the learning system is not biased towards a subset of the variables or categories in the database.

Data format

To analyze the data of N patients (also called 'observations'), each with M different measurements (also called 'variables' or 'dimensions'), e.g., ejection fraction, body mass index (BMI), and image pixels/voxels, by using an ML algorithm, a data matrix X should first be constructed such that the rows of this data matrix correspond to the observations and the columns correspond to the variables (Figure 2). Depending on the database and the problem at hand, X can be either a 'wide' (Figure 2a) or a 'tall' (Figure 2b) data matrix. In the former case, the number of observations is much smaller than the number of variables (N << M; Figure 2c), while in the latter, there is a large group of observations, but each observation has only a few variables (N >> M; Figure 2d).

Generating a data matrix from cardiac images can be performed either on entire images or on selected regions of the images, depending on the learning purpose. When the goal is to use a learning algorithm for modeling the global characteristics of the images, hand-crafted features (e.g., radiomics) extracted based on all the pixels of selected regions of a given image are considered to be the variables of one observation (Figure 2c), which typically leads to a wide data matrix. On the contrary, to model regional image characteristics, however, a region of interest (ROI) or patch consisting of a small group of pixels, thus more easily yielding a tall data matrix (Figure 2d). Further, a series of techniques (e.g., convolution, max pooling or patch-based methods) can further be applied in reducing the size of the image-data matrix to highly informative elements, which often results in a tall/thin matrix, either for classification or for pattern recognition task to identify key regions of interest (7, 25). For example, a data matrix generated by using, say, a single frame from longitudinal, 4 chamber or short-axis view of the heart that is 512x512 pixels in size for a total of N patients would result in creating a N x (512^2) size matrix for input into deep learning algorithms.

Data preparation

To analyze cardiac images in a ML framework, some preprocessing stages are usually carried out. The irrelevant areas of the images can be removed in a 'cropping' stage to focus on learning from useful regions and to prevent learning from extraneous regions (which can also contribute to leakage, as discussed below). If the images that are acquired from a group of subjects have different sizes, they typically should be 'resized' first (26) to a reference image size to construct a data matrix with the same number of variables. More advanced techniques from computational atlases are also necessary to align the anatomy-based data of each subject to a common geometry and temporal dynamics, as spatiotemporal misalignment of input images will increase variance in the input of neurons of a neural network thereby slowing its ability to learn relevant features (27, 28). Another common preprocessing stage is 'noise removal', which helps a learning algorithm to better model the essential characteristics of the images. When the acquired images have poor contrast, a 'histogram equalization' (26) technique can be used to adjust the intensities of the pixels and to increase the contrast of a low contrast region, thus facilitating its interpretation and analysis. The pixel intensities can also be manually adjusted during image acquisition. An example is the changing of the dynamic range of echocardiographic images by an operator. Techniques may also be applied to correct the differences in slice thickness, grey level distribution or even image resolution and different imaging protocols (contrast / non contrast; low dose / high dose). Deep learning techniques, which are robust with regards to image quality, may need to be employed. For CT and MRI images a particular window and level settings may be applied before the deep learning training and normalized to the full intensity ranges.

In cardiovascular imaging, data samples are often represented in 3D and 4D formats which currently may present challenges for the deep learning techniques due to the constraints posed by computing resources related to image sampling as well as large number of input variables. Nevertheless, efficient deep learning techniques have been developed for video analysis and these can be potentially adapted to the direct interrogation of 3D or 4D data often seen in echography, nuclear cardiology or CT (29). An alternative approach to deal with complex multidimensional data is to provide an intermediate simplified image representation. In cardiac imaging, often a bull's eye representation (aka 'polar map') is used for the projection of 3D or 4D image data into a simple 2D (or 2D+time) format. For example, a bull's-eye representation approach has successfully been implemented for deep learning of nuclear cardiology studies (6, 30, 31). It allows data normalization from multiple scans and disparate sources such as motion/thickening, perfusion or flow. Such approaches could also be applied to other cardiovascular modalities.

Thus, data preparation step aids in normalizing and compressing images, with respect to their intensities, anatomical representation and viewpoint etc., across a given study prior to their entry into the ML framework. This process ensures that the algorithm spends more time and capacity learning the features that are important, rather than trying to rectify issues related to intensities or noise levels in the image data. Size of the images can also be reduced in order decrease memory requirements.

Feature engineering and learning

The next stage after data preparation is extracting a set of 'features' from the data matrix to be later used as the input to the learning methods. Feature extraction helps to overcome the following two main problems that can limit the efficient performance of a learning framework: **(i) Curse of dimensionality**: When the data matrix is wide, the variable/feature space of the data can be referred to as 'high-dimensional'. This may lead to an algorithm which fails to learn essential characteristics of the data due to its complexity and poor generalization power when dealing with unseen data — a phenomenon that is referred to as the 'curse of dimensionality' (32, 33). To tackle these problems, the number of observations should increase significantly with the data dimensionality. For example, in pattern recognition, a typical rule of thumb is that there should be at least 5 training examples for each uncorrelated dimension in the representation (34). Moreover, it has been previously suggested that the sample to feature ratio should be between 5-10 depending upon the complexity of the classifier (35-36). However, a significant increase in the number of observations is not always possible, especially for medical data/studies, given that it necessitates the collection of data from a large group of patients. This curse of dimensionality is one of the main reasons why having a large database is desirable to build an efficient learning algorithm.

(ii) Correlated variables: When a database includes correlated variables, a subset of the variables that are mutually uncorrelated may be sufficient to learn the data characteristics effectively (32). Indeed, adding correlated variables to a database will only bring redundant information and does not help the learning algorithm to achieve a better understanding of the data. For the imaging data for example, neighboring pixels typically have similar values and are highly correlated (37).

Given the curse of dimensionality, learning process of algorithms cannot work effectively in data with too many features. Techniques to reduce the number of variables while retaining the most relevant information are critically important; a process called 'feature extraction' and 'dimensionality reduction'. It can be performed either manually using expert knowledge or by algorithms such as principal component analysis or multifactor dimensionality reduction (32–34). The result of the feature extraction process should be a compact set of (ideally uncorrelated) features or variables in the form of a tall matrix that encodes the essential characteristics of the data.

The available approaches for extracting features from the image data can be divided into the following three broad categories (Figure 3): (i) handcrafted methods (e.g., local binary patterns (LBP) (38) and scale invariant feature transform (SIFT) (39)), (ii) classic ML methods for dimensionality reduction (e.g., PCA (40), independent component analysis (ICA) (41), or ISOMAP (42)) and (iii) deep learning methods (43). The methods in the first two categories are manually designed to extract specific types of features from the data, while in the last one, the features are learned from the database itself. Moreover, in the case of deep-learning, techniques such as max-pooling provide effective ways of down-sampling the image size in each layer, eliminating the need for feature extraction and improving overall training performance (25). Nevertheless, the classical feature learning algorithms have some limitations in the data modeling approaches like linearity, sparsity, or lack of hierarchical representation. The deep learning techniques, on the other hand, can learn complex features from the data at multiple levels and do not have limitations of the classical algorithms. However, they need a large-scale database to achieve efficient learning of the data characteristics. To train a deep learning algorithm with a smaller database, the following two main strategies can be used: (i) data augmentation (e.g., by using different types of data/image transformations) (44) and (ii) transfer learning, which works by fine-tuning a deep network that has been pretrained with a different large database (e.g., natural images) (44, 45). However, the field of transfer learning for medical imaging data is currently at its early stages, and more critical insights into its actual relevance will be known in the coming years.

Variable normalization

For a database that is composed of several variables of different nature (e.g., anthropometric or imaging-derived measurements), the values of the variables lie in different ranges. Direct usage of these variables may bias the learning system towards the characteristics of the variables with larger values despite the usefulness of the variables with smaller values in solving a given problem. To deal with such challenges, a ‘variable normalization’ approach can be used to transform the variables such that they all lie in the same range prior to entering the learning phase (33, 34). Variable normalization is especially helpful for a deep learning algorithm, as it helps achieve faster convergence of a deep neural network (46).

Missing variable estimation

Machine learning algorithms often need complete datasets. If data are missing, the options are to exclude those subjects, encode them as missing or to impute missing values (45, 47, 48). In cardiovascular imaging, 2D images are normally collected from multiple views, e.g., for volumetric measurements, and 3D images are composed of multiple 2D slices. These images can also be acquired throughout the cardiac cycle. When some of the 2D views are not accessible or when a group of 2D images at some points during the cardiac cycle or in a 3D volume are artefactual/missing, an imputation technique can estimate these images or the parameters extracted from them (48). Thanks to development of the new deep learning algorithms, such as generative adversarial networks (GAN) (49), missing images can often be estimated (50) based on the available data, although the physiological relevance of their content is not fully guaranteed. However, it should be acknowledged that most of the imputation methods assume that the missing observations occur at random, are missing completely at random, or are missing not at random (51).

Researchers should consider whether the missing observations carry any specific biases (e.g., selection bias or immortal time bias). While there is no clear guidance or cutoff on what proportion of missing data warrants the use of data imputation techniques (52) and given that this answer depends on the complexity of the addressed problem, missing value imputation is best utilized whenever possible as it provides evidence about the robustness of the learned models regardless of its impact on model performance (53).

Feature selection

An important phase in designing a classic ML system is to determine the optimal number of preserved features. This determination can be performed by using a 'feature selection' technique where a larger than required set of features is first extracted and then a subset with discriminative information is selected (34, 54). When a deep learning algorithm is used, the optimal features are automatically learned during the end-to-end training of the algorithm, and utilizing an independent feature selection method is often not required (43, 55**).**

Outliers

An observation is considered as an outlier if its values deviate substantially (or significantly, if looked through a statistical test) from the average values of a database, which may be attributed to measurement error, variability in the measurement, or abnormalities due to disease (33, 34). Even though outliers can negatively influence and mislead the training process resulting in less accurate models, they may carry relevant information related to the given task. Thus, outliers should be carefully examined to see if this comes from improper measurements that should be repeated, or not. With the existence of outliers, learning algorithms and/or a performance metrics that are robust to outliers can be useful alternatives. Methods robust to outliers including but not limited to decision trees and k-nearest neighbor (KNN) should be employed as much as possible (33). If no other solution exits, the removal of outliers, using an outlier detection approach (56), may be considered and the selection criterion along with the proportion of samples removed should be reported.

Class imbalance

A significant imbalance in data classes (e.g., healthy vs. diseased) is quite common in medical datasets because, on the one hand, the majority of subjects in a database are usually healthy and, on the other hand, because collecting patient data for some rare diseases is difficult and is not always possible. As a result, the performance of the learning algorithm might be skewed, as it only learns the characteristics of the larger sized categories. This problem is referred to as ‘class imbalance’ and can be dealt with in the following three established ways: (i) rebalancing the categories using ‘under-sampling’ or ‘over-sampling’ (i.e., making the different classes similarly sized by omitting samples from the larger class or by up-sampling the data in the smaller class), (ii) giving more importance (i.e., weight) to the samples of smaller categories during the learning process (34), and iii) utilizing synthetic data generation methods, such as the synthetic minority over-sampling technique (SMOTE) (57). While the random over-sampling generates new data by duplicating some of the original samples of the minority class or category, SMOTE interpolates values using a k-nearest neighbor technique to synthesize new data instances (58) . Recent advances in deep generative techniques, such as GAN or variational autoencoders or the use of loss functions that are robust to data imbalance (59), have made it possible to tackle complicated imbalanced data based on the learning strategies.

Data shift

Data shift is a common problem that afflicts the ML models in cardiovascular imaging in which the distribution of the database that is used for testing the performance of the learning models or systems may differ from the distribution of the training data. This may occur when the data acquisition conditions or the systems that are used for collecting the test data change from when the training dataset was acquired, and could induce i) a covariate shift – a shift in the distribution in the covariates, ii) a prior probability shift – a difference in the distribution of the target variable, or iii) a domain shift – a change in measurement systems or methods. It is imperative to assess and treat the shifts that may occur in the dataset prior to evaluating a model (60).

Data leakage

Data leakage is a major problem in ML, in which data outside of the training set seeps into the model while building the model. This event could lead to error-prone or invalid ML models. Data leakage could occur if the same patient's data is used in the training and testing sets and is generally a problem in complex datasets, such as time series, audio and images, or graph problems.

Strategic steps for developing the checklist:

- The data format for training a machine learning algorithm should be large and the ratio of the observations/measurements (i.e., N/M) should be at least five.
- When the data matrix is wide, a feature extraction/learning algorithm or dimensionality reduction technique should be used.
- Redundant features should be removed, and variables should be normalized.
- Outliers should be addressed/removed
- Missing features should be imputed using relevant methods.
- Dataset shift, leakage and class imbalance are common pitfalls and should be evaluated.

3. Selection of Machine Learning Models

Model selection is the process of identifying the model that yields the best resolution and generalizability for the project and can be defined at multiple levels, i.e., learning methods, algorithms, and tuning hyperparameters. Learning methods include supervised, unsupervised, and reinforcement learning. Importantly, supervised learning is a method that learns from labeled data, i.e., data with outcome information to develop a prediction model, while unsupervised learning aims to find patterns and association rules in data that do not have labels (Figure 4).

Common algorithms, such as regression or instance-based learning, often handle high-dimensional data well and tend to perform better or equivalent to complex algorithms on small datasets while retaining the interpretability of the model. To achieve better performance, simple algorithms, or weak learners, may be combined in various ways using ensemble methods, such as boosting, bagging, and stacking, which sacrifice the interpretability. More complex algorithms that are also difficult to interpret, including neural networks, can outperform simpler models given an adequate amount of data. A subset of neural networks, known as deep convolutional neural networks (7, 25), are particularly useful for finding patterns in image data without the need for feature extraction (61-64). The implementation of an algorithm can vary significantly in terms of the size and complexity (e.g., the size and number of features in a random forest decision tree, the number and complexity of kernels applied in an SVM, and the number and type of nodes and layers in a neural network) of the algorithms.

Regardless of the choice of the algorithms, it is imperative to perform hyperparameter tuning and model regularization to produce the optimal performance (65, 66). These processes may be more important than selecting the types of algorithms that could impact the interpretability, simplicity, and accuracy. When performing model selection (especially involving large-scale data and/or deep learning methods), hardware constraints (e.g., memory size, cache, parallelism etc.) are often a key limiting factor beyond model performance. However, recent advances towards developing hardware accelerators (e.g., graphical processing units, tensor processing units) and growing convenience/abstraction by cloud computing could improve the overall process of model training/selection/optimization. Moreover, selecting the best model often involves the relative comparison of performance between different models. Therefore, the purposeful selection of loss function and the metric that represents it (e.g., absolute error, mean-squared error etc.), which in turn is heavily influenced by the model choice, dataset, and particular problem/task to solve, becomes fundamental in selecting an appropriate model.

The size and complexity of algorithms should be chosen carefully to minimize the bias, the model error on the training dataset, and the variance, the model error on the validation dataset. Simpler models may underfit the data; they may generalize better (lower variance) at the cost of lower accuracy (higher bias). Further, overfitting (high variance and low bias) may come from a too complex model or insufficient representative training samples. Several considerations including size, complexity/dimensionality, number of features and nonlinear relationships among variables in the dataset guides the choice of the initial algorithm and its complexity, but the final algorithm design (including the choice of hyperparameters) is determined empirically or by specific optimization and cross-validation.

Finally, an essential factor in algorithm selection is the need for the interpretability of the model's decisions, i.e., an understanding of which input features caused the model to make the decision it made. Interpretability may be extremely important for certain learning tasks and less important for others. Regression, decision trees, and instance-based learning methods are generally highly interpretable, while methods to interpret the function of deep neural networks are still evolving; saliency mapping, class activation, and attention mapping are some example methods for neural network interpretation and visualization (67,68). New mechanisms for understanding the workings of machine learning models (69-72), and approaches for probabilistic deep learning (73,74) further provide an opportunity to develop models to balance between both inference and prediction.

Strategic steps for developing the checklist:

- For the initial model development, always select the simplest algorithm that is appropriate for the available data.
- The size of the dataset and the complexity of the employed algorithm should be considered to achieve a good compromise between 'bias' and 'variance' in the estimations.
- Complex algorithms must be benchmarked to the performance of the initial simple model across several metrics.
- Tune the hyperparameters to optimize the models and to increase performance.

4. Model Assessment

The next step after selecting a learning model is to evaluate the generalizability by applying it to new data, i.e., assessment of its performance on unseen data. Ideally, model assessment should be performed by randomly dividing the dataset into a ‘training set’ for learning the data characteristics, a ‘validation set’ for tuning the hyperparameters of the learning model, and a ‘test set’ for estimating its generalization error, where all the three sets have the same probability distribution (i.e., the statistical characteristics of the data in these three sets are identical). However, in many domains, including cardiovascular imaging, having access to a large dataset is often difficult, thus preventing model assessment using three independent data subsets. As mentioned in the previous section, the ratio of the training samples to the number of measured variables should be at least five to ten (34, 35), depending on the dataset and complexity of the classifier (75), to learn the data characteristics properly. If this criterion is not met, the data are called scarce. In this situation, the data may be divided into two subsets for training and final validation of the learning algorithm. However, the results may depend on the random selection of the samples. Therefore, the training set can then be further partitioned into two subsets, but this process is repeated several times by selecting different training and testing subjects to obtain a good estimate of the generalization performance of the learning algorithm (33). This method of model assessment can be performed via ‘cross-validation’ or ‘bootstrapping’, as further explained below. These techniques ensure that (i) the learning model is trained properly given that the majority of the data samples can be used in the training process, (ii) the learning model is not biased towards the characteristics of a subset of the data and (iii) the optimal values of the hyperparameters of the learning model (e.g., the number of layers in a neural network and the neurons in each layer) can be determined (33).

Cross-validation

This technique works by dividing the data into multiple nonoverlapping training and testing subsets (also called folds) and using the majority of the folds for training a learning model and the remaining folds for evaluating its performance (32–34). The cross-validation process can be implemented in one of the following ways.

i. **k-fold cross-validation**: The data is randomly partitioned into k folds of roughly equal sizes, and in each round of the cross-validation process, one of the folds is used for testing the learning algorithm and the rest of the folds are used for its training (Figure 5). This process is repeated k times such that all folds are used in the testing phase and the average performance on the testing folds is computed as an unbiased estimate of the overall performance of the algorithm (32, 33).

ii. **Leave-one-out cross-validation**: In this technique, the number of the folds is equal to the number of the observations in the database, and in each round, only one observation is used for testing the learning algorithm.

iii. **Monte-Carlo cross-validation**: In this method of cross-validation, there is no limit to the number of the folds, and a database can be randomly partitioned into multiple training and testing sets. The training samples are randomly selected 'without replacement', and the remaining samples are used for the testing group (Figure 6I) (76).

Bootstrapping

This method works by randomly sampling observations from a database 'with replacement' to form a training set whose size is equal to the original database. As a result, some of the observations can appear several times in the training set, while some may never be selected. The latter observations are called 'out-of-bag' and are used to test the learning algorithm. This process is repeated multiple times to estimate the learning method's generalization performance (Figure 6 II) (33, 76). While bootstrapping tends to drastically reduce the variance, it often tends to provide more biased results, more importantly when dealing with small sample sizes.

Strategic steps for developing the checklist:

- Model assessment should be performed by randomly dividing the dataset into training, validation and testing data when applicable.
- When data is inadequate or scarce, model assessments using cross-validation and/or bootstrapping techniques should be performed to obtain a good estimate of the generalization performance of the learning algorithm.
- Typical numbers for *k* in a *k*-fold cross-validation should be 5 and 10.
- Consider using leave-one-out cross validation as an appropriate choice when the data is small.

5. Model Evaluation

The reporting of accuracy in ML is closely linked to the reporting of summary statistics, and the same background and assumptions apply. While a review of statistical theory is out of scope for the PRIME Checklist, we encourage the readers to obtain a clear understanding of the statistics for classification and prediction (77-83). Most of the following section applies to supervised learning algorithms, for which labels are used in the definition of the performance measures. Unsupervised learning is more difficult to evaluate but should also evaluate the relevance of the output data representation and the stability of the results against the data and model parameters.

For classification tasks, the accuracy is the percentage of data that is correctly classified by the model, which could be influenced by the quality of the expert annotations. The balance of classes in the training data is also a known source of bias. As such, a prerequisite for reporting accuracy measures is to provide a clear description of the data material used for training and validation. We further suggest balancing the class data according to prevalence when possible, or that balanced accuracy measures are reported (84).

The model parameters (e.g., initialization scheme, number of feature maps, and loss function), regularization strategies (e.g., smoothness and dropouts), and hyperparameters (e.g., optimizer, learning rate, and stopping criterion) also play a part in the model performance. A second prerequisite is, therefore, to provide a clear description of how the ML model was generated. We further suggest that the certainty of the accuracy measure is reported where applicable, for instance, by estimating the ensemble average and variance from several models generated with random initialization. Additionally, cross-validation analysis should be added to underline the robustness of the model, especially for limited training and test data (see the previous section). Furthermore, to assess the generalizability of the algorithm, it is necessary to report the accuracy of the model by testing the data from different geographical locations with similar statistical properties and distributions (85).

A report of the accuracy for ML algorithms in cardiovascular imaging will depend on the method and problem. For instance, the classification of disease from image features differs from the classification of image pixels in semantic segmentation, both in terms of the measures reported and of the risk in use.

For multiclass/label classification, we suggest using a statistical language close to the clinical standard. For instance, the report sensitivity, specificity, and odds ratio should be used instead of the precision, recall, and F1 score. This will also ensure that true negative outcomes are considered (86). Nonetheless, for classification tasks, the confusion matrix should normally be included but could be supplementary material. For image segmentation problems, we suggest reporting several measures to summarize both the global and local deviations, such as the mean absolute error (MAE), the Dice score to summarize the average performance and the Hausdorff distance metric to capture local outliers.

When the output of the regression or segmentation algorithms are linked to clinical measurements (e.g., ejection fraction), we suggest Bland-Altman plots as for conventional evaluation of the image measurements, and we stress the importance of comparing the performance with several expert observers for both intra- and inter- expert variability.

For the classification of disease from image features, the cost of misclassification should be clearly conveyed, e.g., rare diseases may not be properly represented in the dataset. The balance of classes should reflect the prevalence of the disease of interest, and scoring rules based on estimated probability distributions should be used for the accuracy reporting when possible, instead of direct classification. The choice of the scoring rule used for the decision, e.g., mean squared error, Brier score, and log-loss, should be rationalized. The common classification scores (sensitivity, specificity, positive- and negative predictive value) should include a full ROC analysis to provide a more in-depth evaluation of the detection performance. It is also relevant to include benchmark results from alternative ML methods as well as more traditional techniques, such as logistic regression.

Strategic steps for developing the checklist:

- Use a statistical language close to the clinical standard and introduce new measures only when needed.
- Balance the classes according to prevalence where available or report balanced accuracy measures.
- Estimate the accuracy certainty, e.g., from an ensemble of models, to strengthen the confidence in the values reported.
- Include Bland-Altman plots when machine learning is linked to clinical measurements.
- Include an inter-/ intra-observer variability measures as a reference where possible.
- The risk of misclassification should be conveyed, and appropriate scoring rules for decisions may be needed for the classification of a disease.

6. Best Practices for Model Replicability

The reproducibility of scientific results is essential to make progress in cardiac medicine. The ability to reproduce findings helps to ensure the validity and correctness, as well as enabling others to translate the results into clinical practice. However, there are several complementary definitions of reproducibility. We focus here primarily on technical replicability (87); i.e., the ability to independently confirm published results of a model by inspecting and executing data and code under identical conditions. Technical replicability is especially important in ML projects, which often involve custom software scripts, the use of external libraries, and intensive or expensive computation. Actions taken at any point in an ML workflow, from quality control and data preparation into suitable data structures to algorithm development to the visualization of results, are often based upon heuristic judgments, and there are potentially numerous justifiable analytic options. Ultimately, these selections may significantly alter the results and conclusions.

The first step for making ML projects reproducible could be the release of all the original code written for a project. There are several options for the publication of code. When possible, we suggest uploading source code with software and packages' version information as supplementary material alongside the manuscript. Other options include permanent archival on a lab website, or per-project archival on open source and public source code repositories if permitted by investigator's institution. Manuscripts should explicitly state where and how the code may be downloaded and under what license.

Although there are numerous open source licenses available, in most cases, either of two licenses will suffice: the Massachusetts Institute of Technology/Berkeley Software Distribution (MIT/BSD) licenses (https://opensource.org/licenses/MIT) (the MIT and BSD licenses are essentially equivalent) and the GNU General Public License (GNU GPLv3). The MIT/BSD licenses allow published code to be distributed, modified, and executed freely without liability or warranty; the GNU GPLv3 license allows the same with the additional restriction that all software-based upon the original code must also be freely available under the GNU GPLv3 license, meaning others cannot reuse the original code in a closed-source product.

Although the availability of code is required for technical replicability, equally important is the availability of the data used in the project (86). Clinical data should be anonymized, or if anonymization is not possible (as in the case of some genetic data), then data should be made available to other researchers with appropriate IRB approval. Other options include the generation of synthetic datasets with the same statistical properties as the original dataset, a field of study called differential privacy. Manuscripts must state where both the raw and manipulated/transformed data may be obtained and justify any restrictions to data availability. All data should also be accompanied by a codebook (also known as a data dictionary) containing clear and succinct explanations of all variables and class labels along with detailed description of their data types and dimensions.

Finally, we note that even in the case of freely available data and open-source code, it can be difficult to reproduce the results of published work due to the complexities of software package versioning and interactions between different computing environments. We, thus, suggest that authors make the entire analyses automatically reproducible through the use of software environments (e.g., Docker containers, https://www.docker.com/). Analyses in software containers may be freely downloaded and run from beginning to end by other scientists, greatly

Strategic steps for developing the checklist:

- Release the code and upload data as supplementary information alongside the manuscript when possible for non-commercial use; otherwise, consider making the code and data available via an academic website for non-commercial use as permissible.
- Use the MIT/BSD or GPLv3 license to release open-source code.
- Release a codebook (data dictionary) with clear and succinct explanations of all variables.
- Document the exact version of all external libraries and software environments.
- Consider the use of Docker containers or similar packaging environments such as Sphinx for straightforward technical reproducibility and to generate reliable code/software manuals.

improving the technical replicability. Moreover, documentation generation tools (e.g., Sphinx, https://www.sphinx-doc.org/en/master/) or easy-to-launch demos (e.g. through Jupiter python notebooks) should be employed.

7. Reporting Limitations, Biases and Alternatives

"All models are wrong, but some are useful" is a well-known statistical aphorism attributed to George Box. Accurate reporting and acknowledgement of limitations are required for manuscripts incorporating ML (ML). Any statistical model or ML algorithm incorporates some assumptions regarding the data. All model assumptions should be affirmatively identified and checked with the dataset utilized in the manuscript, and the results should be reported in the manuscript or supplementary material. The algorithms used in computational research efforts span a large spectrum of complexity. Generally, more basic models and algorithms should first be investigated before additional complexity is incorporated into models or different algorithms are selected. Deep learning models should be benchmarked against simpler models whenever possible, especially when applied to tabular data. Statistical or ML models incorporating large numbers of variables (e.g., polygenic risk score models) should be benchmarked against standard clinical risk prediction models using more traditional clinical variables.

Concordant findings from multiple, independent datasets dramatically increase the scientific value of manuscripts, since it decreases the likelihood that the algorithms have been erroneously overfitted to the idiosyncratic features of a certain dataset. Deep learning models are especially notorious for harnessing spurious or confounding features of the dataset to perform well. For example, Zech and Badgeley et al. reported a case where a convolutional neural network trained on a health system's chest X-rays used the presence of a "PORTABLE" label on X-ray images to predict cardiomegaly with high accuracy (88). Furthermore, in the case of supervised ML involving human-annotated variables or outcomes, it should be noted that ML algorithms will recapitulate the underlying biases of the humans who constructed the dataset.

Strategic steps for developing the checklist:

- Affirmatively identify and check relevant model assumptions and report the findings.
- Benchmark complex algorithms against simpler algorithms and justify the use of more complex models.
- Benchmark algorithms, incorporating high-dimensional data or novel data sources, against standard clinical risk prediction models

Summary and Future Directions

As artificial intelligence and ML technologies continue to grow, three specific areas of opportunities will need further consideration for future standardization. First, there has been growing enthusiasm in the use of automated machine learning (auto-ML) platforms that democratize machine-learning strategies. Second, using the 'multiomics-approach', clinical and other data like smart-phones based health-data could be integrated with imaging variables to provide more algorithmic sophistication and objectivity to the existing taxonomy of risk factors and cardiac diseases (89-91). Finally, sophisticated algorithms and variations of GAN will be increasingly used to synthesize data that closely resemble the distribution of the input data (92-94). This approach may be particularly fruitful for the field of simulation and in-silico clinical trials, which were recently recognized by the Food and Drug Administration (FDA) as key new directions to validate novel devices and therapies (95). In this context, recent studies have combined computational modeling with ML for synthetic data generation or tracking a disease course (96). Moreover, the ML research presents unprecedented opportunities for restructuring the industry, research, and medical alliance. This fact is well recognized by the FDA, which has mandated the standardization and applications of ML software as medical devices (97). With the advancement of organizations and cardiac medicine and imaging towards the actualization of precision medicine, the PRIME Checklist would need to be updated continuously as ML algorithms continue to transform cardiovascular imaging practice over the next decade.

Acknowledgement: We thank Jai K. Nahar, MD, MBA and James E. Tcheng, MD for reviewing and providing constructive feedback for the document.

References:

1. Douglas PS, Cerqueira MD, Berman DS et al. The Future of Cardiac Imaging: Report of a Think Tank Convened by the American College of Cardiology. JACC Cardiovasc Imaging 2016;9:1211-1223.
2. Sengupta PP, Shrestha S. Machine Learning for Data-Driven Discovery: The Rise and Relevance. JACC Cardiovasc Imaging 2019;12:690-692.
3. Dey D, Slomka PJ, Leeson P et al. Artificial Intelligence in Cardiovascular Imaging: JACC State-of-the-Art Review. J Am Coll Cardiol 2019;73:1317-1335.
4. Winther HB, Hundt C, Schmidt B et al. ν-net: Deep Learning for Generalized Biventricular Mass and Function Parameters Using Multicenter Cardiac MRI Data. JACC Cardiovasc Imaging 2018;11:1036-1038.
5. Tan LK, McLaughlin RA, Lim E, Abdul Aziz YF, Liew YM. Fully automated segmentation of the left ventricle in cine cardiac MRI using neural network regression. J Magn Reson Imaging 2018;48:140-152.
6. Betancur J, Commandeur F, Motlagh M et al. Deep Learning for Prediction of Obstructive Disease From Fast Myocardial Perfusion SPECT: A Multicenter Study. JACC Cardiovasc Imaging 2018;11:1654-1663.
7. Madani A, Arnaout R, Mofrad M. Fast and accurate view classification of echocardiograms using deep learning. NPJ Digit Med 2018;1.
8. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. J Am Coll Cardiol 2017;69:2657-2664.
9. Zheng T, Xie W, Xu L et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 2017;97:120-127.

10. Dawes TJW, de Marvao A, Shi W et al. Machine Learning of Three-dimensional Right Ventricular Motion Enables Outcome Prediction in Pulmonary Hypertension: A Cardiac MR Imaging Study. Radiology 2017;283:381-390.

11. Lang RM, Badano LP, Mor-Avi V et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr 2015;28:1-39.e14.

12. Zhang J, Gajjala S, Agrawal P et al. Fully Automated Echocardiogram Interpretation in Clinical Practice. Circulation 2018;138:1623-1635.

13. Ouyang D, He B, Ghorbani A et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020;580:252-256.

14. Fahmy AS, Rausch J, Neisius U et al. Automated Cardiac MR Scar Quantification in Hypertrophic Cardiomyopathy Using Deep Convolutional Neural Networks. JACC Cardiovasc Imaging 2018;11:1917-1918.

15. Bai W, Sinclair M, Tarroni G et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J Cardiovasc Magn Reson 2018;20:65.

16. Johnson KW, Torres Soto J, Glicksberg BS et al. Artificial Intelligence in Cardiology. J Am Coll Cardiol 2018;71:2668-2679.

17. Al'Aref SJ, Anchouche K, Singh G et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. Eur Heart J 2019;40:1975-1986.

18. Ghojogh B, Crowley M. The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial. https://arxiv.org/abs/1905.12787: arXiv, 2019.

19. Dhurandhar A, Shanmugam K, Luss R, Olsen PA. Improving Simple Models with Confidence Profiles. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). Montréal, Canada.: Neural Information Processing Systems (NIPS), 2018.

20. Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide-prediction, machine learning and causal inference. Int J Epidemiol 2019: 1-7.

21. Leng S, Xu Z, Ma H. Reconstructing directional causal networks with random forest: Causality meeting machine learning. Chaos 2019;29:093130.

22. Vandenbroucke JP, von Elm E, Altman DG et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Int J Surg 2014;12:1500-24.

23. Cohen O, Malka O, Ringel Z. Learning Curves for Deep Neural Networks: A Gaussian Field Theory Perspective. https://arxiv.org/abs/1906.05301: arXiv, 2019.

24. Mehta P, Wang CH, Day AGR et al. A high-bias, low-variance introduction to Machine Learning for physicists. Phys Rep 2019;810:1-124.

25. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical Image Analysis using Convolutional Neural Networks: A Review. J Med Syst 2018;42:226.

26. Gonzalez RC, Woods RE. Digital Image Processing. 3rd ed: Pearson, 2007.

27. Duchateau N, Craene MD, Pennec X, Merino B, Sitges M, Bijnens B. Which reorientation framework for the atlas-based comparison of motion from cardiac image sequences? In: Durrleman S., Fletcher T., Gerig G., Niethammer M. (eds) Spatio-temporal Image Analysis for Longitudinal and Time-Series Image Data. STIA 2012. Lecture Notes in Computer Science, vol 7570.: Springer, Berlin, Heidelberg, 2012:25–37.

28. Duchateau N, De Craene M, Piella G et al. A spatiotemporal statistical atlas of motion for the quantification of abnormal myocardial tissue velocities. Med Image Anal 2011;15:316-28.

29. Ullah A, Ahmad J, Muhammad K, Sajjad M, Baik SW. Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. IEEE access 2018;6:1155-1166.

30. Betancur J, Hu LH, Commandeur F et al. Deep Learning Analysis of Upright-Supine High-Efficiency SPECT Myocardial Perfusion Imaging for Prediction of Obstructive Coronary Artery Disease: A Multicenter Study. J Nucl Med 2019;60:664-670.

31. Spier N, Nekolla S, Rupprecht C, Mustafa M, Navab N, Baust M. Classification of Polar Maps from Cardiac Perfusion Imaging with Graph-Convolutional Neural Networks. Sci Rep 2019;9:7569.

32. Bishop CM. Pattern recognition and machine learning: Springer-Verlag New York, 2006.

33. Hastie T, Tibshirani, Robert., J. F. The Elements of Statistical LearningData Mining, Inference, and Prediction. Second ed. New York: Springer New York, 2009.

34. Koutroumbas K, Theodoridis S. Pattern recognition. 4th ed: Academic Press, 2008.

35. Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 2003;19:1484-91.

36. Dernoncourt D, Hanczar B, Zucker J-D. Analysis of feature selection stability on high dimension and small sample data. Computational Statistics and Data Analysis 2014;71:681-693.

37. Hyvärinen A, Hurri J, Hoyer PO. Natural Image Statistics: A Probabilistic Approach to Early Computational Vision. 1 ed: Springer-Verlag London, 2009.

38. Ojala T, Pietikäinen MK, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 2002;24:971–987.

39. Lowe DG. Object recognition from local scale-invariant features. ICCV '99: Proceedings of the International Conference on Computer Vision 1999;2:1150–1157.

40. Jolliffe IT. Principal Component Analysis. 2nd ed: Springer-Verlag New York, 2002.

41. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw 2000;13:411-30.

42. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290:2319-23.

43. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.

44. Shin HC, Roth HR, Gao M et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging 2016;35:1285-98.

45. Bengio Y. Deep Learning of Representations for Unsupervised and Transfer Learning. JMLR: Workshop and Conference Proceedings 27;: PMLR, 2012:17-37.

46. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. https://arxiv.org/abs/1502.03167: arXiv, 2015.

47. Troyanskaya O, Cantor M, Sherlock G et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17:520-5.

48. Tabassian M, Alessandrini M, Jasaityte R, Marchi LD, Masetti G, D'hooge J. Handling missing strain (rate) curves using K-nearest neighbor imputation. Tours, France: 2016 IEEE International Ultrasonics Symposium (IUS), 2016:1-4.

49. Goodfellow I, Pouget-Abadie J, Mirza M et al. Generative Adversarial Nets. Advances in Neural Information Processing Systems 27 2014;2:2672–2680.

50. Shang C, Palmer A, Sun J, Chen KS, Lu J, Bi J. VIGAN: Missing View Imputation with Generative Adversarial Networks. Proc IEEE Int Conf Big Data;2017:766-775.

51. Rhodes W. Improving Disparity Research by Imputing Missing Data in Health Care Records. Health Serv Res 2015;50:939-45.

52. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinical Epidemiology 2019:63-73.

53. Liu Y, Gopalakrishnan V. An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data. Data (Basel) 2017;2:1-23.

54. Inza I, Larrañaga P, Blanco R, Cerrolaza AJ. Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med 2004;31:91-103.

55. Arnaout R, Curran L, Chinn E, Zhao Y, Moon-Grady A. Deep-learning models improve on community-level diagnosis for common congenital heart disease lesions. https://arxiv.org/abs/1905.12787: arXiv, 2018.

56. Atkinson A, Riani M. Robust Diagnostic Regression Analysis. 1 ed: Springer-Verlag New York, 2000.

57. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 2002;16:321–357.

58. Last F, Douzas G, Bacao F. Oversampling for Imbalanced Learning Based on K-Means and SMOTE. 12Dec2017 ed. https://arxiv.org/abs/1711.00837: arXiv, 2017.

59. Rezende DJ, Mohamed S. Variational Inference with Normalizing Flows. 32nd Int. Conf. Mach. Learn. ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France: JMLR: W&CP, 2015:1530–1538.

60. Subbaswamy A, Schulam P, Saria S. Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport. Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS). Naha, Okinawa, Japan: PMLR, 2019:3118-3127.

61. Wang J, Ding H, Bidgoli FA et al. Detecting Cardiovascular Disease from Mammograms With Deep Learning. IEEE Trans Med Imaging 2017;36:1172-1181.

62. Litjens G, Ciompi F, Wolterink JM et al. State-of-the-Art Deep Learning in Cardiovascular Image Analysis. JACC Cardiovasc Imaging 2019;12:1549-1565.

63. Retson TA, Besser AH, Sall S, Golden D, Hsiao A. Machine Learning and Deep Neural Networks in Thoracic and Cardiovascular Imaging. J Thorac Imaging 2019;34:192-201.

64. Avendi MR, Kheradvar A, Jafarkhani H. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. Med Image Anal 2016;30:108-119.

65. Kostoglou K, Robertson AD, MacIntosh BJ, Mitsis GD. A Novel Framework for Estimating Time-Varying Multivariate Autoregressive Models and Application to Cardiovascular Responses to Acute Exercise. IEEE Trans Biomed Eng 2019;66:3257-3266.

66. Al'Aref SJ, Singh G, van Rosendael AR et al. Determinants of In-Hospital Mortality After Percutaneous Coronary Intervention: A Machine Learning Approach. J Am Heart Assoc 2019;8:e011160.

67. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining Explanations: An Overview of Interpretability of Machine Learning 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) 2018:80-89.

68. Fong RC, Vedaldi A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. 2017 IEEE International Conference on Computer Vision (ICCV) 2017:3449-3457.

69. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. https://arxiv.org/abs/1702.08608: arXiv, 2017.

70. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 2018:1-42.

71. Olah C, Satyanarayan A, Johnson I et al. The Building Blocks of Interpretability. Distill, 2018. Available at: https://distill.pub/2018/building-blocks/

72. Samek W, Wiegand T, Müller K-R. Explainable Artificial Intelligence: Understanding,

Visualizing and Interpreting Deep Learning Models. https://arxiv.org/abs/1708.08296: arxiv, 2017.

73. Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science 2015;350:1332-8.

74. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Proceedings of the 33 rd International Conference on Machine Learning. New York, NY, USA: JMLR: W&CP, 2016.

75. Raudys Š. Statistical and neural classifiers: an integrated approach to design: Springer, 2001.

76. Kuhn M, Johnson K. Applied Predictive Modeling: Springer, New York, NY, 2013.

77. Wheelan C. Naked statistics. stripping the dread from the data. 1st ed: W.W. Norton & Company, 2013:302.

78. Mlodinow L. The Drunkard's Walk: How Randomness Rules Our Lives: Vintage, 2009.

79. Wasserman L. All of statistics : a concise course in statistical inference. 1 ed: Springer-Verlag New York, 2004.

80. Urdan TC. Statistics in Plain English. 2nd ed: Psychology Press, 2005.

81. Cohen PR. Empirical methods for artificial intelligence: MIT Press, 1995.

82. Box GEP, Hunter JS, Hunter WG. Statistics for Experimenters: Design, Innovation, and Discovery. 2nd ed: Wiley, 2005.

83. Sabo R, Boone E. Statistical Research Methods: A Guide for Non-Statisticians. 1 ed: Springer-Verlag New York, 2013.

84. Wainer J, Franceschinell RA. An empirical evaluation of imbalanced data strategies from a practitioner's point of view. https://arxiv.org/abs/1810.07168: arXiv, 2018.

85. Abazeed M. Walking the tightrope of artificial intelligence guidelines in clinical practice. The Lancet Digital Health, 2019:PE100.

86. Tharwat A. Classification assessment methods. Applied Computing and Informatics: Elsevier BV, 2018.

87. McDermott MBA, Wang S, Marinsek N, Ranganath R, Ghassemi M, Foschini L. Reproducibility in Machine Learning for Health. Presented at the ICLR 2019 Reproducibility in Machine Learning Workshop: https://arxiv.org/abs/1907.01463: arXiv, 2019.

88. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15:e1002683.

89. Natarajan S, Jain A, Krishnan R, Rogye A, Sivaprasad S. Diagnostic Accuracy of Community-Based Diabetic Retinopathy Screening With an Offline Artificial Intelligence System on a Smartphone. JAMA Ophthalmol 2019;137:1182-1188.

90. Rajalakshmi R, Subashini R, Anjana RM, Mohan V. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. 2018;32:1138–1144.

91. Wasserlauf J, You C, Patel R, Valys A, Albert D, Passman R. Smartwatch Performance for the Detection and Quantification of Atrial Fibrillation. Circ Arrhythm Electrophysiol 2019;12:e006834.

92. Suarez J. GAN You Do the GAN GAN? https://arxiv.org/abs/1904.00724: arXiv, 2019.

93. Wang X, He K, Hopcroft JE. AT-GAN: A Generative Attack Model for Adversarial Transferring on Generative Adversarial Nets. https://arxiv.org/abs/1904.07793v3: arXiv, 2019.

94. Chang C-H, Yu C-H, Chen S-Y, Chang EY. KG-GAN: Knowledge-Guided Generative Adversarial Networks. https://arxiv.org/abs/1905.12261: arXiv, 2019.

95. Morrison T. How Simulation Can Transform Regulatory Pathways. FDA, 2018. Available at: https://www.fda.gov/science-research/about-science-research-fda/how-simulation-can-transform-regulatory-pathways. Accessed October 14, 2019.

96. Kagiyama N, Shrestha S, Farjo PD, Sengupta PP. Artificial Intelligence: Practical Primer for Clinical Research in Cardiovascular Disease. J Am Heart Assoc 2019;8:e012788.

97. Software as a Medical Device (SaMD): Key Definitions. International Medical Device Regulatory Forum, 2013. Available at: https://www.fda.gov/medical-devices/digital-health/software-medical-device-samd. Accessed October 14, 2019.

Figure Legends

Figure 1: Machine learning pipeline
Schematic diagram of a general ML pipeline. The data section consists of project planning, data collection, cleaning, and exploration. The modelling section describes the model building, in which hyperparameter tuning and the dimensionality reduction process, such as feature selection and engineering, model optimization and selection, and evaluation, are included. Finally, the reporting segment consists of the reporting mechanisms of the analysis, including reproducibility and maintenance, and a description of the limitations and alternatives.

Figure 2: Schematic demonstration of short/wide (a) and tall/thin (b) data matrices and the way that they can be created from the image data.
In a short and wide data matrix, the number of observations is much smaller than the number of variables (N<<M). Considering different regions of interest (ROI) on the whole image of a given patient, the extraction of hand crafted features (e.g., radiomics features) may lead to a short and wide data matrix, as the number of features extracted for each ROI per image is typically larger than the number of samples or patients (c). To make a tall and thin data matrix from the image data, an image can be divided to many (overlapping) ROIs or patches, each with a small number of pixels (d). The extraction of pixel data as features from each patch per patient may be much smaller in size than the total number of patients.

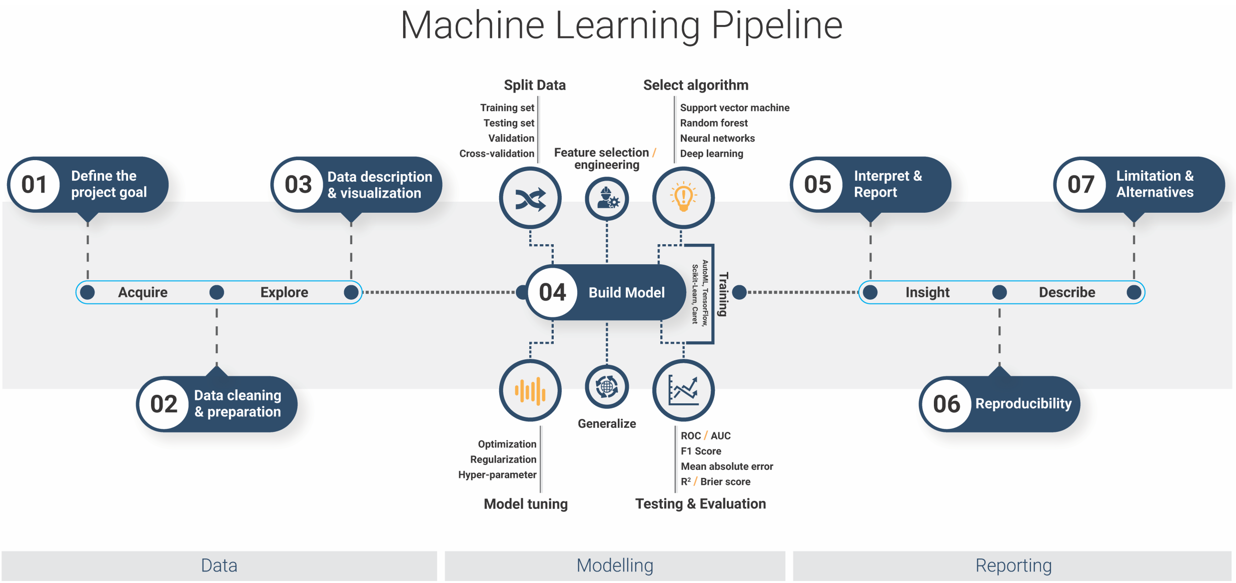
Figure 3: The main approaches for feature engineering and learning.
The hand-engineering approaches are manually designed to extract certain types of features from the data; for example, Local Binary Pattern (LBP) and Scale-Invariant Feature Transform (SIFT) derives the properties from the image such as object recognition or edge detection. The classic learning techniques use data samples to learn their characteristics for dimensionality reduction, but they have limitations in their data modeling techniques, such as linearity, sparsity or lack of hierarchical representation. Principal component analysis (PCA) applies orthogonal transformation to produce linear combination of uncorrelated variables that best explains the variability of the data whereas independent component analysis (ICA) transforms the dataset into independent components to reduce dimensionality. Deep learning methods, however, can learn complex features from the data at multiple levels in various hidden layers.

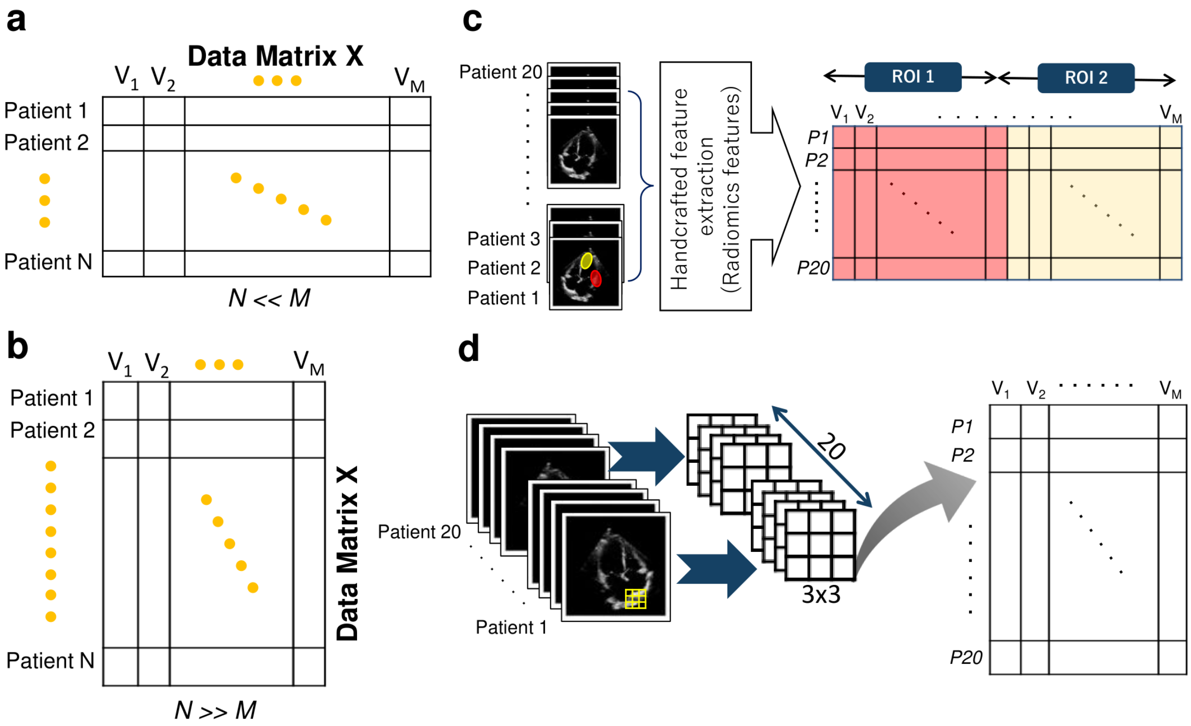
Figure 4: Model selection process
Illustration of the model selection process, which consists of identifying the two classes of ML. A) In Supervised learning method, after the hyperparameter tuning, data can be applied to two different tasks: classification or regression, depending on the type of the outcome. If the outcome is a category, then classification can be performed whereas if the outcome variable is a numeric value, regression may be applied for prediction. B) Unsupervised learning method, the data in which the data are either utilized for clustering, topical modeling, or representing the data distribution while reducing the dimensionality of the data according to the problem to be solved.

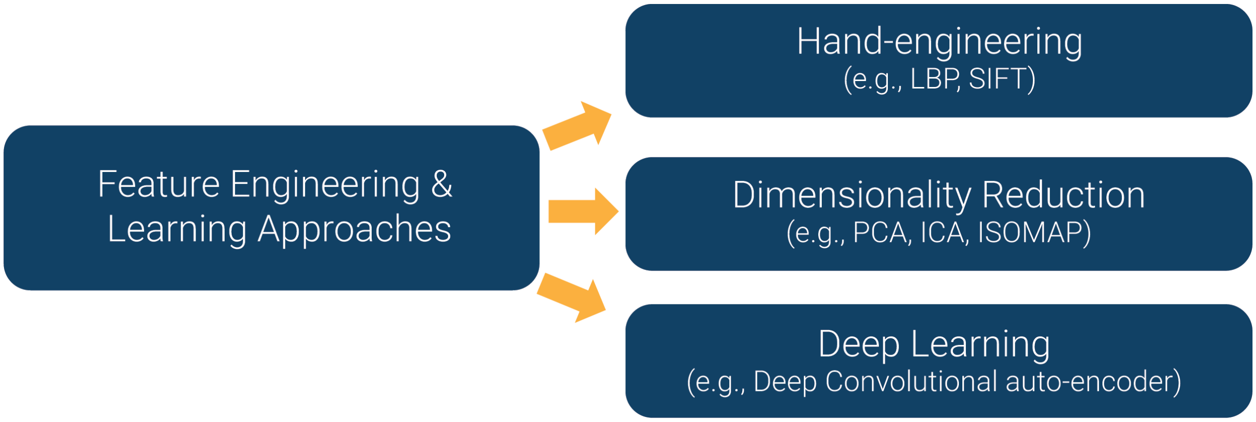
Figure 5: **Schematic illustration of the k-fold cross-validation process.** Data are randomly partitioned into k distinct folds, and in each round, (k-1) folds are used for training the learning algorithm, and the kth fold is used for testing its performance. This process is repeated k times such that all folds are used in the testing phase.

Figure 6: Schematic illustration of the monte-carlo and bootstrap resampling methods. I) Monte-Carlo cross-validation performed in k rounds. In each round, the training and testing samples are randomly selected without replacement from the original data. II) The bootstrapping process, which can be performed in B rounds. In each round, the training data is generated by randomly sampling from the original data with replacement. The samples that are not included in the training dataset (i.e., out-of-bag samples) form the testing dataset.

Central Illustration: Steps for building a machine learning pipeline and the reporting items in a checklist. This illustration provides the principal requirements for building a checklist (at every step of the model building process) to enable the precise application of predictive modeling, consistent reporting of model specifications and results in the field of cardiovascular imaging. CV = cross validation; GPL = general public license; LOOCV = leave one out cross validation; ML = machine learning; S/W = software.


Machine Learning Pipeline
01 Define the project goal
02 Data cleaning & preparation
03 Data description & visualization
04 Build Model
05 Interpret & Report
06 Reproducibility
07 Limitation & Alternatives
Acquire
Explore
Insight
Describe
Split Data
Training set
Testing set
Validation
Cross-validation
Feature selection / engineering
Select algorithm
Support vector machine
Random forest
Neural networks
Deep learning
Training
AutoML, TensorFlow, Scikit-Learn, Caret
Generalize
Optimization
Regularization
Hyper-parameter
Model tuning
ROC / AUC
F1 Score
Mean absolute error
R² / Brier score
Testing & Evaluation
Data
Modelling
Reporting


a
Data Matrix X
V1 V2 VM
Patient 1
Patient 2
Patient N
N << M
b
V1 V2 VM
Patient 1
Patient 2
Patient N
Data Matrix X
N >> M
c
Patient 20
Patient 3
Patient 2
Patient 1
Handcrafted feature extraction (Radiomics features)
ROI 1
ROI 2
V1 V2 VM
P1
P2
P20
d
Patient 20
Patient 1
20
3x3
V1 V2 VM
P1
P2
P20

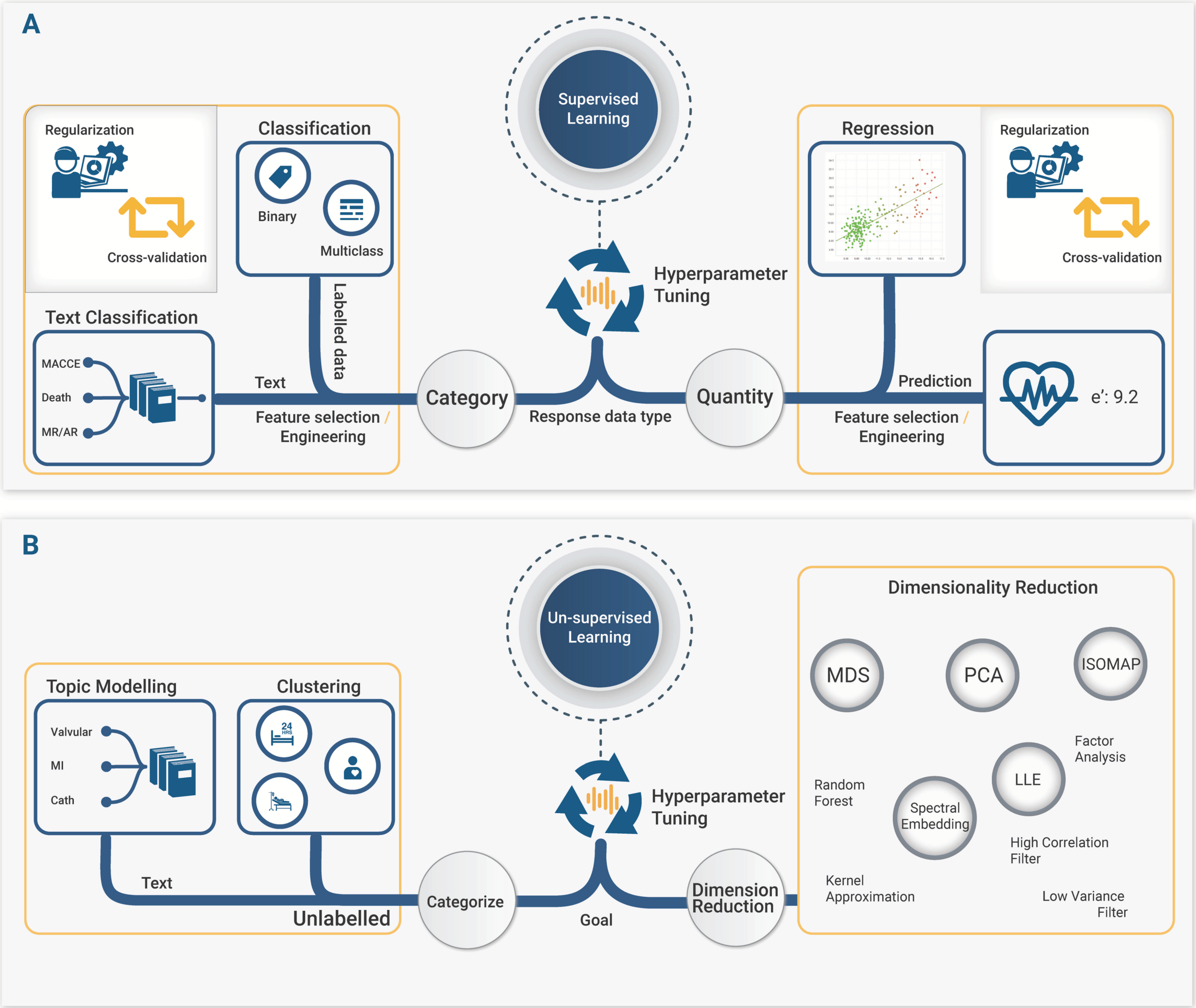


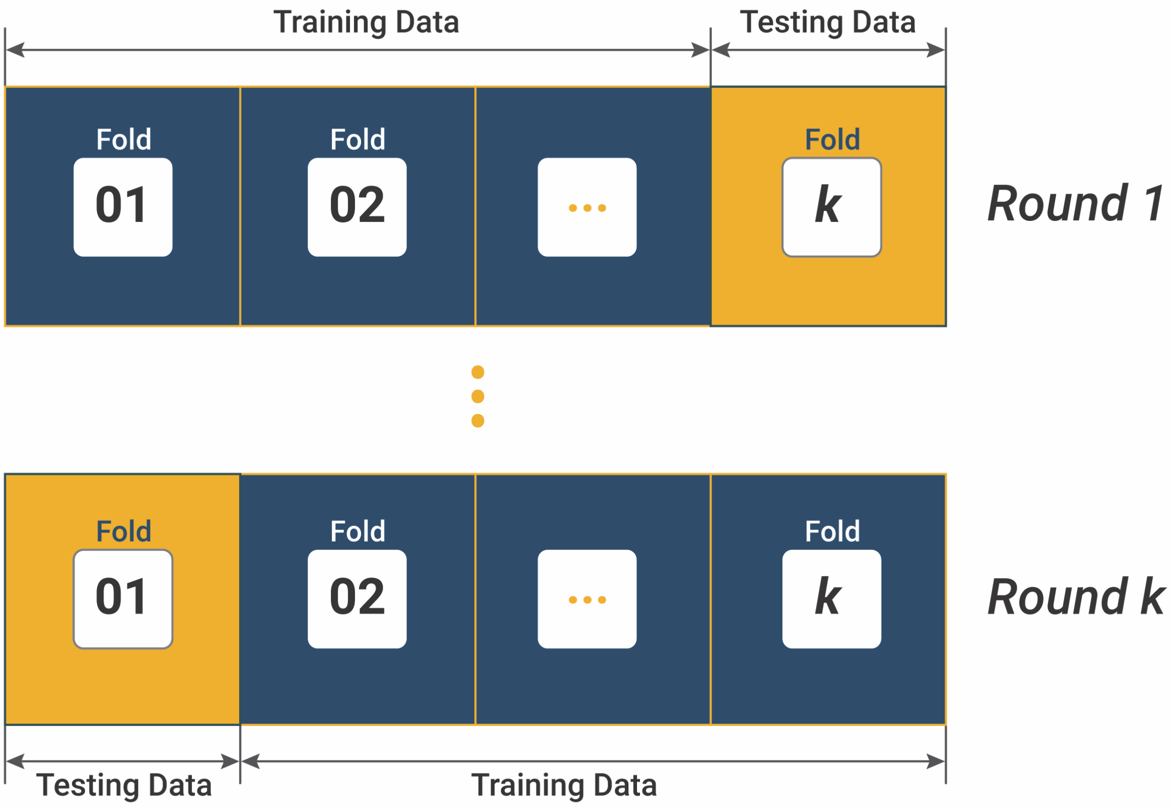
Feature Engineering & Learning Approaches

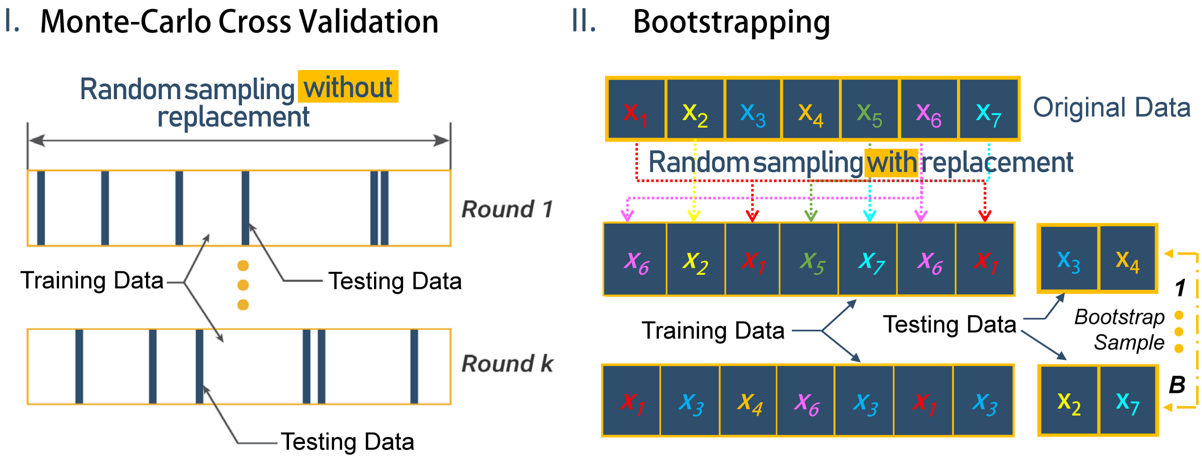
Hand-engineering (e.g., LBP, SIFT)

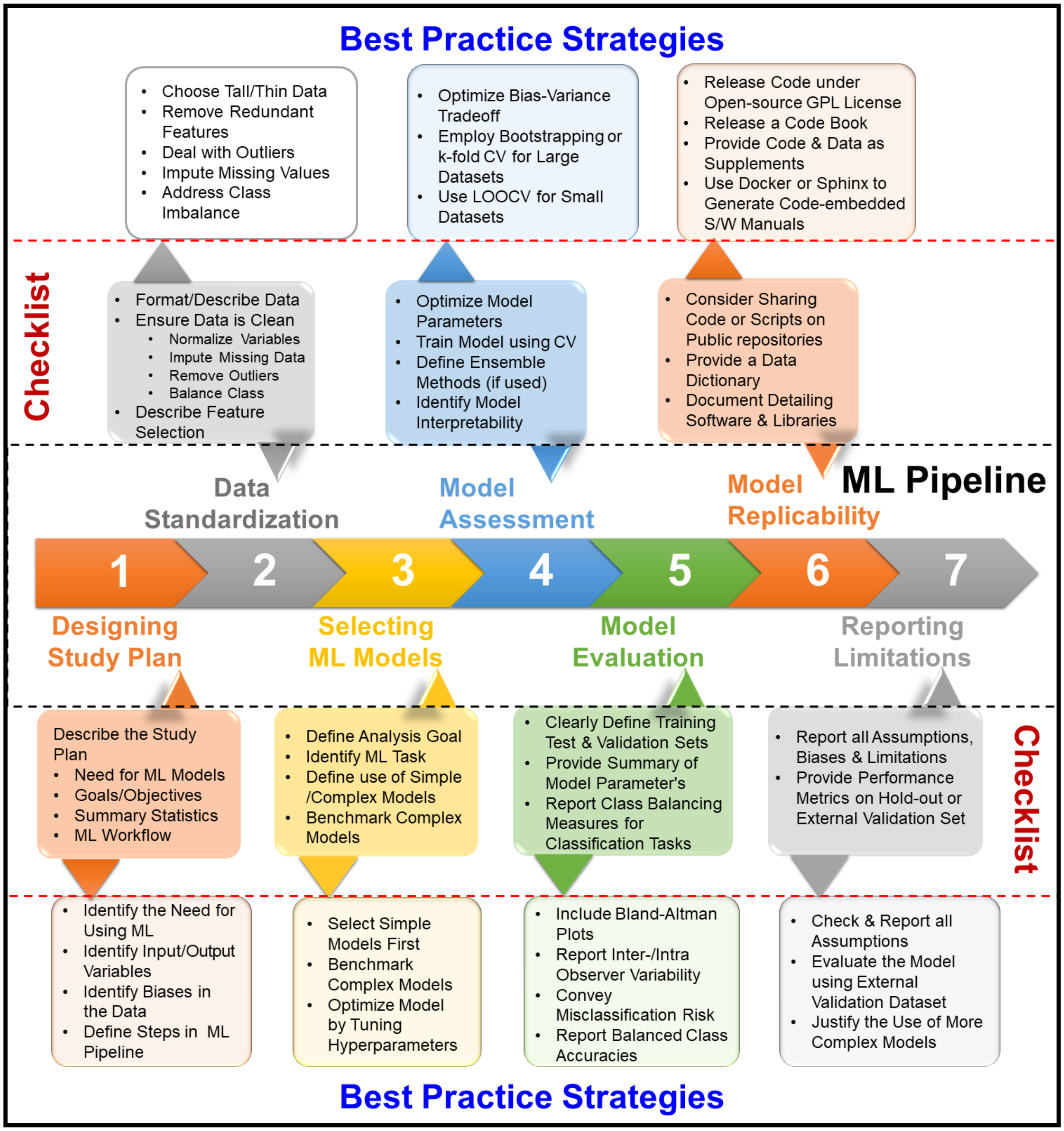
Dimensionality Reduction (e.g., PCA, ICA, ISOMAP)

Deep Learning (e.g., Deep Convolutional auto-encoder)




Training Data
Testing Data
Fold
01
Fold
02
...
Fold
k
Round 1
Fold
01
Fold
02
...
Fold
k
Round k
Testing Data
Training Data


I. Monte-Carlo Cross Validation
Random sampling without replacement
Round 1
Training Data
Testing Data
Round k
Testing Data
II. Bootstrapping
x_1 x_2 x_3 x_4 x_5 x_6 x_7
Original Data
Random sampling with replacement
x_6 x_2 x_1 x_5 x_7 x_6 x_1
x_3 x_4
1
Training Data
Testing Data
Bootstrap Sample
x_1 x_3 x_4 x_6 x_3 x_1 x_3
x_2 x_7
B



Best Practice Strategies

- Choose Tall/Thin Data
- Remove Redundant Features
- Deal with Outliers
- Impute Missing Values
- Address Class Imbalance

- Optimize Bias-Variance Tradeoff
- Employ Bootstrapping or k-fold CV for Large Datasets
- Use LOOCV for Small Datasets

- Release Code under Open-source GPL License
- Release a Code Book
- Provide Code & Data as Supplements
- Use Docker or Sphinx to Generate Code-embedded S/W Manuals

Checklist

- Format/Describe Data
- Ensure Data is Clean
  - Normalize Variables
  - Impute Missing Data
  - Remove Outliers
  - Balance Class
- Describe Feature Selection

- Optimize Model Parameters
- Train Model using CV
- Define Ensemble Methods (if used)
- Identify Model Interpretability

- Consider Sharing Code or Scripts on Public repositories
- Provide a Data Dictionary
- Document Detailing Software & Libraries

Data Standardization

Model Assessment

Model Replicability

ML Pipeline

1 2 3 4 5 6 7

Designing Study Plan

Selecting ML Models

Model Evaluation

Reporting Limitations

Describe the Study Plan
- Need for ML Models
- Goals/Objectives
- Summary Statistics
- ML Workflow

- Define Analysis Goal
- Identify ML Task
- Define use of Simple /Complex Models
- Benchmark Complex Models

- Clearly Define Training Test & Validation Sets
- Provide Summary of Model Parameter's
- Report Class Balancing Measures for Classification Tasks

- Report all Assumptions, Biases & Limitations
- Provide Performance Metrics on Hold-out or External Validation Set

Checklist

- Identify the Need for Using ML
- Identify Input/Output Variables
- Identify Biases in the Data
- Define Steps in ML Pipeline

- Select Simple Models First
- Benchmark Complex Models
- Optimize Model by Tuning Hyperparameters

- Include Bland-Altman Plots
- Report Inter-/Intra Observer Variability
- Convey Misclassification Risk
- Report Balanced Class Accuracies

- Check & Report all Assumptions
- Evaluate the Model using External Validation Dataset
- Justify the Use of More Complex Models

Best Practice Strategies

Table 1. Checklist for Standardized Reporting of Machine Learning Investigations

Section	Checklist item	Page #
1	**Designing the Study Plan**	
1.1	Describe the need for the application of machine learning to the dataset	
1.2	Describe the objectives of the machine learning analysis	
1.3	Define the study plan	
1.4	Describe the summary statistics of baseline data	
1.5	Describe the overall steps of the machine learning workflow	
2	**Data Standardization, Feature Engineering, and Learning**	
2.1	Describe how the data were processed in order to make it clean, uniform, and consistent	
2.2	Describe whether variables were normalized and if so, how this was done	
2.3	Provide details on the fraction of missing values (if any) and imputation methods	
2.4	Describe any feature selection processes applied	
2.5	Identify and describe the process to handle outliers, if any	
2.6	Describe whether class imbalance existed and which method was applied to deal with it	
3	**Selection of Machine Learning Models**	
3.1	Explicitly define the goal of the analysis e.g., regression, classification, clustering	
3.2	Identify the proper learning method used (e.g., supervised, reinforcement learning etc.) to address the problem	
3.3	Provide explicit details on the use of simpler, complex, or ensemble models	
3.4	Provide the comparison of complex models against simpler models if possible	
3.5	Define ensemble methods, if used	
3.6	Provide details on whether the model is interpretable	
4	**Model Assessment**	
4.1	Provide a clear description of data used for training, validation, and testing	
4. 2	Describe how the model parameters were optimized (e.g., optimization technique, number of model parameters etc.)	
5	**Model Evaluation**	
5.1	Provide the metric(s) used to evaluate the performance of the model	
5.2	Define the prevalence of disease and the choice of the scoring rule used	
5.3	Report any methods used to balance the numbers of subjects in each class	
5.4	Discuss the risk associated to misclassification	
6	**Best Practices for Model Replicability**	
6.1	Consider sharing code or scripts on a public repository with appropriate copyright protection steps for further development and non-commercial use	
6.2	Release a data dictionary with appropriate explanation of the variables	
6.3	Document the version of all software and external libraries used	
7.	**Reporting Limitations, Biases and Alternatives**	
7.1	Identify and report the relevant model assumptions and findings	
7.2	If well performing models were tested on a hold-out validation dataset, detail the data of that validation set with the same rigor as that of training dataset (see section 2 above)	

Introduction

Artificial Intelligence	Branch of computer science concerned with building intelligent systems that automatically behave based on the data and experience they learn. Typically, there are two types of AI: general and applied AI. General AI is a self-sufficient system that can possess a cognition capability comparable to, or even surpassing, that of humans. Applied AI is a functional system that is specialized for a purpose. Natural language processing and ML are forms of AI.
Machine Learning	A subfield of applied AI that concerns algorithms, statistical modeling, and data analysis and that learns from the data to detect patterns and make assessments with minimal human intervention. There are 3 main classes to apply ML: supervised, unsupervised, and re-enforcement learning. Semi-supervised learning is also considered a class of ML.

Designing the study plan

Model Interpretability	A degree to which an individual can comprehend the reason and decision made by the algorithm. Higher the interpretability of the ML model, easier it is for an individual to comprehend the reason to a decision reached by an algorithm.
Model Complexity	Refers to the number of features/covariates/variables and the transformations and linearity (or non-linearity) of the features that are included in the predictive model. It can also refer to the learning and computational complexity of a given learning algorithm.
Bayesian inference	Process to interpret probability and represent the confidence given an occurrence of an event. It is based on Bayes' statistical method that assigns probabilities to events or parameters based on the current data but updates as the probabilities as more data is obtained.
Systematic error	The errors that are produced from measurement errors. This may occur if the measurement unit is faulty, incorrectly used, or the data incorrectly interpreted or entered.

Coverage error	The bias that is introduced in the data when all components of the population are not adequately covered. It may occur if a component is included more than actual estimate from the survey (over-coverage), failure to include adequately (under-coverage), or misclassified.
Selection bias	The bias in the data that occurs when the individual or group of data for analysis is not collected with proper randomization, and, therefore, is not representative enough of the population of interest.
Hyperparameters	Parameters that are used to modulate how the model learns from the data and typically tuned before the learning process begins, or during the cross-validation process.
Neural Networks	ML models that are inspired from the neurons of the human brain. A network has input layers and output layers and may have one or more intermediate layers. Some networks have only one or a few layers, such as perceptron, or multiple layer perceptron, which only feeds the data forward. Deep learning relies on a network with several layers in the architecture and has a backward data feed to minimize error.
Deep neural network / Deep Learning	A type of ML based on neural networks made of several (in general, at least 3) intermediate layers. It is used for supervised and unsupervised tasks but requires large amount of data compared to traditional shallow learners. It is also known as deep neural network. Convolutional neural network, generative adversarial network, recurrent neural network are some examples popular in computer vision and medical imaging.
Independent and Identical Distribution	If the data or a random variable is collected from the same probability distribution as all others and is mutually independent such that the outcome of the data is not dependent on the previous observation.

Data standardization, feature engineering and learning

Matrix	Mathematical object that contains numeric values, symbols, or expressions arranged in a rectangular array of columns and rows.
Histogram equalization	A method in image processing to adjust the contrast of the image using the image’s histogram.
Pixel intensity	The value of a pixel’s brightness. In normal medical images, pixel intensities range from 0 to 255 where greater values indicate brighter colors.

Data Annotation	A process of labeling the data available in various formats (e.g., image, video, audio or text) to make it usable for ML.
Data augmentation	A technique to add data or variability to the data.
Transfer learning	Transfer learning is a method of ML in which the knowledge gained from another pretrained model is applied to learn from another set of data. This method can reduce the need for data by orders of magnitude.
Dimensionality Reduction	An unsupervised ML technique to find a new representation of the data while reducing the features that describe the data. This method is useful for avoiding overfitting and producing simpler models.
Multicollinearity	An issue in statistics and ML regression models in which two or more variables provide the same information due to their close relationship.
Generative Adversarial Network (GAN)	Generative models using deep neural network architecture that consists of the following two stages: the generator stage, which generates new examples, and the discriminator stage, which classifies if the generated examples are real (with the same distribution as the problem domain) or fake. Importantly, generative models became popular due to their ability to generate highly realistic images (in terms of appearance, not necessarily physiological content).
Hand-crafted features	In image-based classification tasks, "hand-crafted" features refer to properties derived using certain manually predefined algorithm based on the expert knowledge using the information present in the image itself. Some examples of hand-crafted features include local binary patterns, scale invariant feature transform and histogram of oriented gradients etc.
Variational Autoencoders	In deep learning, variational autoencoders (VAE) are a class of neural networks that can learn to compress data (i.e., image, text or sequence) and perform dimensionality reduction completely in an unsupervised manner. Instead of letting a neural network learn from an arbitrary function, VAEs learn from the parameters of a probability distribution by modeling input data.
Linearity	In statistical terms, linearity means that the response variable is a linear combination of the parameters of independent variables (i.e., regression coefficients) and the predictor/dependent variables.

Sparsity	In numerical terms, sparsity is estimated as the number of zero-valued elements divided by the total number of elements in a data array or matrix. Input data will be considered sparse when most of the elements (> 50% of data) of a data array or matrix are zero i.e., when its sparsity is greater than 0.5.

Selection of Machine Learning Models

Supervised	A class of ML that learns from labeled data to predict or classify the outcome of interest. The most common supervised learning tasks are regression and classification.
Unsupervised	A class of ML that learns from unlabeled data. The most common unsupervised learning tasks are clustering and dimensionality reduction.
Semi-supervised Learning	A class of ML that learns from labeled and unlabeled data in the dataset. Typically, the proportion of unlabeled data is higher than the proportion of labeled data.
Overfitting	A case when the model fits the training data too closely, but the generalization of the model is unreliable.
Underfitting	A case when the model neither fits the training data nor generalizes to new data. An underfit model results in low generalization and unreliable predictions.
Clustering	An unsupervised ML task in which the fractions of data with some notion of similarity are grouped together while keeping the others separate. K-means and agglomerative hierarchical clustering are popular clustering algorithms.
Ensemble learning	An ML process to combine multiple models to obtain better predictive performance compared to the performance from a single model. Bagging, boosting, and stacking are three main methods of ensemble learning.
Bagging	Bagging also known as ‘bootstrap aggregating’ is an ensemble technique that is designed to improve the stability and the accuracy of ML algorithms. The objective of bagging is simple, generate several random training sets and “average” their predictions in order to obtain a model with a lower variance. Thus, bagging reduces variance and helps to avoid overfitting.

Boosting	In ML boosting is also an ensemble technique that attempts to create a strong learner from a number of weak learners. The main objective of boosting is to increase the complexity of models that suffer from high bias especially in the case of underfitting. In simple terms, boosting tries to reduce the error in predictions and hence reduces bias.
Saliency mapping	Saliency can be seen as a kind of image segmentation. In computer vision, a saliency map is an image that shows each pixel's unique quality. The goal of a saliency map is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze.

Model Assessment

Model	The mathematical representation of the process that is governed by the data. The data is partitioned into training or testing, and the training set is generally used to generate the model that can be tested for its generalizability in ML. For example, a model is presented as $\hat{y} = \beta_1 x_1 + \ \beta_2 x_2 \dots \beta_n x_n$ in linear regression.
Model Parameters	Model parameters are internal configuration variables such as coefficients in linear regression (e.g., β1, β2, etc. in the equation above) that are estimated from the data and describe the model.

Model Evaluation

Classification	A supervised ML technique that attempts to classify data into binary scores or categories.
Regression	A supervised ML technique that attempts to predict a continuous value.

Loss function	A method to measure error made by the model in prediction for a single training example. Mean squared error, likelihood loss, mean absolute error are some of the common loss functions in evaluating the model. It is also known as error function.
Regularization	Technique to reduce complexity and avoiding overfitting by adding penalty term which can e.g., shrink or eliminate the coefficients, to smoothen the estimated trend or classification boundary.
Precision	Proportion of samples that were classified or identified as positive. $$\frac{True\ Positive}{True\ Positive + False\ Positive}$$
Recall	Proportion of positive samples classified or identified correctly. $\frac{True\ Positive}{True\ Positive+False\ Negative}$
F1 score	Weighted average of the precision and recall where 1 is the best value and 0 is the worst. $$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$
Brier score	Method to verify the accuracy of a classification where 0 is for 100% accurate and 1 is 100% inaccurate. The most common formulation of the Brier score is $$BS = \frac{1}{N}\sum_{t=1}^{N}(\mathrm{f_t} - o_t)^2$$ where N is the number of instances, f_t is the forecast probability (i.e. 25% chance) and o_t corresponds to the outcome (1 if it happened, 0 if it didn't).

Software Engineering Best Practices and Data Availability for Reproducibility

Source code repository	A central file storage location used by version control system to manage and store several versions of source codes.

Practical illustration of how to apply the checklist developed in the PRIME guidelines

The finalized PRIME reporting guidelines are shown in Table 1. The guidelines checklist is structured to correspond with the various facets of the building machine learning systems (data preprocessing, analysis, model development etc) and is intended to capture critical information to enable correct application of machine learning (ML) models and allow consistent reporting of model results in cardiovascular imaging studies.

Basically, checklist is a set of essential items for authors to consider when reporting results from ML models for publication in cardiovascular imaging. The checklist begins with an item that relates to providing a detailed description of the overall study plan. It is very important to clearly define and establish the need for the application of ML approaches to a given problem, as it becomes the foundation for all the other steps in ML workflow. Model selection, algorithm complexity, performance evaluation and assessment all are interlinked and often depend on the task that one is trying to solve. Similarly, we also recommend listing and reporting items related to limitations, biases and alternative analysis strategies, to satisfy the need for precision, completeness, and transparency in reporting results of modeling studies. Apart from these, we also created separate checklist items for each process in the ML workflow to provide a list of reporting items to be included in a research articles and reports.

To further determine its ease of use, applicability and understand how well it captures all the practical steps of developing predictive models, we gathered information from two recent articles that developed ML models in cardiovascular imaging. The entire manuscripts were searched thoroughly to extract any information reported that was in reference to the PRIME recommendation's checklist items. A complete checklist along with page numbers referring to each item is provided in Table S1 (see below). Both of these studies implemented deep learning models either to improve the diagnosis of congenital heart disease or augment the computer-assisted echocardiographic interpretation. Almost 90% of the items in the PRIME reporting checklist were addressed or satisfied by the two studies considered, indicating that these reports/articles were executed with PRIME-analogous guidelines in mind. Certain missing checklist items such as *"Benchmark complex models against more simplistic models if possible"* were in fact not applicable due to the specific nature of the selected ML model and the nature of the study (e.g., deep learning applied to video or image analysis. Thus, the guidelines listed here could be effective in capturing the reporting requirements of studies employing machine learning approaches in cardiovascular imaging.

Supplemantary Table S1

Section	**Checklist item**	**Madani et al (1)**	**Arnaout R et al (2)**
1	**Designing the study plan**		
1.1	Describe the need for the application of machine learning to the dataset	pg 1	pg 2-3
1.2	Describe the objectives of the machine learning analysis	pg 1	pg 2-3
1.3	Define the study plan	pg 2, Fig 1	pg 4-7
1.4	Describe the summary statistics of baseline data	pg 2, Fig 1	pg 4-7
1.5	Describe the overall steps of machine learning workflow	pg 3, pg 4, Fig 2, Table 1	pg 4,7
2	**Data standardization, feature engineering, and learning**		
2.1	Describe how the data were processed in order to make it clean, uniform, and consistent	Fig 1, Fig 3, Table 1, pg 5, pg 6	pg 4
2.2	Describe whether variables were normalized and if so, how this was done	pg 5-7	pg 5
2.3	Provide details on the fraction of missing values (if any) and imputation methods	pg 5-7	N/A
2.4	Perform and describe feature selection process	pg 5-7	pg 5
2.5	Identify and describe the process to handle outliers, if any	N/A	N/A
2.6	Describe whether class imbalance existed and which method was applied to deal with it	pg 4	pg 5
3	**Selection of Machine Learning Model**		
3.1	Explicitly define the goal of the analysis e.g., regression, classification, clustering	pg 1-7	pg 2-3, 10
3.2	Identify the proper learning method used (e.g., supervised, reinforcement learning etc.) to address the problem	pg 1, pg 5	pg 5
3.3	Provide explicit details on the use of simpler, complex, or ensemble models	pg 4	pg 5-10
3.4	Provide the comparison of complex models against simpler models if possible	pg 2, Fig 5	N/A
3.5	Define ensemble methods, if used	N/A	pg 5-6
3.6	Provide details on whether the model is interpretable	pg 2-7, Fig 6	N/A
4	**Model Assessment**		
4.1	Provide a clear description of data used for training, validation, and testing	pg 7	pg 5-6
4.2	Describe how the model parameters were optimized (e.g., optimization technique, number of model parameters etc.)	pg 7	pg 5
5	**Model Evaluation**		
5.1	Provide the metric(s) used to evaluate the performance of the model	pg 4-6, Figs 4-6	pg 7-9
5.2	Define the prevalence of disease and the choice of the scoring rule used	X	X
5.3	Report any methods used to balance the numbers of subjects in each class	pg 5	pg 5
5.4	Discuss the risk associated to misclassification	pg 6	X
6	**Best Practices for Model Replicability**		
6.1	Consider sharing code or scripts on public repository with appropriate copyright protection steps for further development and non-commercial use	pg 5-7	pg 5,7

6.2	Release data dictionary with appropriate explanation of the variables	pg 5-7	N/A
6.3	Document version of all software and external libraries	pg 5-7	pg 5
7	**Reporting limitations, biases and alternatives**		
7.1	Identify and report the relevant model assumptions and findings	X	X
7.2	If well performing models were tested on a hold-out validation dataset, detail the data of that validation set with the same rigor as that of training dataset (see section 2 above)	X	X

References:

1. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. NPJ Digit Med. 2018; 1.

2. **Arnaout R, Curran L, Chinn E, Zhao Y, Moon-Grady A.** Deep-learning models improve on community-level diagnosis for common congenital heart disease lesions. **https://arxiv.org/abs/1809.06993**