A Primer on Machine Learning and Artificial Intelligence

Winok Lapidaire, Maryam Alsharqi, Andrew Fletcher, Paul Leeson

Artificial intelligence (AI) is a broad, non-technical term referring to intelligence exhibited by machines. Its applications range from facial recognition software, self-driving cars to fraud detection. A subfield of AI known as machine learning (ML) allows machines to learn progressively, without continuous human input on how to learn. It can identify patterns and learn rules for a specific goal, based on datasets that the machine has been provided with.[1] ML methods developed for application within medicine are being used in clinical practice at an increasingly faster rate. To be able to use ML wisely in cardiovascular medicine, a thorough understanding of the methodological strengths and limitations, as well as the opportunities and risks of ML clinical applications is required. This primer provides an overview of ML algorithms used in clinical cardiovascular applications, ML algorithm requirements and processes, and examples of ML applications in cardiovascular medicine and their integration into clinical care.

Machine learning terminology and techniques

Supervised learning

Machine learning algorithms can be broadly divided into supervised learning, unsupervised learning, and reinforcement learning. Supervised learning algorithms use a dataset with known true values (labels) to learn to predict or classify a particular characteristic of interest. Labels can be categorical (*e.g.*, ischemic, dilated, and hypertrophic cardiomyopathies), or continuous (*e.g.*, a chamber volume). Ideally this is an undisputed fact, gold-standard measurement or other "ground-truth," rather than an indirect estimate or assumption. The algorithm learns underlying patterns to link them with the known labels,[2-4] so that this can be applied to estimate the values for the characteristic of interest in unlabeled "testing

data." Supervised learning algorithms can be in form of support vector machines, decision trees, and random forests. Support vector machines are a type of supervised algorithm that assigns datapoints to categories predefined by the labelled input data. It first maps examples from the labelled training data in a Oxford way that maximizes the distance between categories. For linear classification the training data space is two-dimensional, but for non-linear classification the training data space can be multidimensional.[5] To estimate the outcome categories in unlabeled testing data, it assigns the testing data to the category that it lies relatively closest to in the training data space. Decision trees have internal nodes that split into branches or edges representing decisions towards a predicted outcome. When data goes through the tree, its path depends on the data values. Different combinations of values lead to a different path and different predicted outcome. The decisions are set up in a way where the labelled training input most often leads to the correct label in the output. Random forests methods create a multitude of decision trees. The data goes through all trees and each tree provides a class prediction. The model's prediction is the outcome that is predicted by the most trees. Having multiple decision trees, trained on different parts of the same training data, reduces the variance, and improves performance.[4]

Unsupervised learning

In unsupervised learning, the training dataset is not labeled, and the goal is to identify patterns in a dataset. Unsupervised learning algorithms can be in form of dimensionality reduction or clustering methods. Dimensionality reduction techniques, such as principal component analysis, are used to reduce the number of input variables whilst keeping as much of the variance contained in the raw data as possible. A principal component analysis is a relatively simple dimensionality reduction tool. Dimensionality reduction is often performed before cluster analysis. Clustering methods, where training data is grouped by its similarity on a set of input parameters, are another form of unsupervised learning. However, clustering can also be performed in a supervised manner if it is based on known (labeled) groups. A combination of models can combine the benefits of supervised and unsupervised learning and reduce overfitting (ensemble learning). This can be done in different ways. In fact, different sets of training data can be created on which the same type of model is trained in parallel and weighed equally (bagging). This minimizes the variance (*e.g.*, a random forest is an ensemble of individual decision trees). Many different types of models can be trained sequentially on the same dataset (boosting). Each model is weighted according to their performance and corrects the errors of the previous models. This reduces bias. Alternatively, several different models deliver their outputs to one final model which determines which model produces the lowest error (stacking).[6] Reinforcement learning processes dynamic data and learns a set of rules by a process of trial and error. The data changes continuously and the algorithm deals with sequential decisions aiming to achieve a particular outcome.[6]

Deep learning

Deep learning refers to a collection of machine learning algorithms that can combine raw inputs into layers of intermediate features. Deep learning methods can combine supervised

and unsupervised learning. When sufficient labeled data is available, features tuned to a specific problem can be combined into a predictor. Deep neural networks are inspired by human brains. They have a series of hidden layers between the input (values in the dataset) and the output (predicted outcomes). Each layer is a mathematical manipulation that feeds into the next layer.[7] The weights between nodes in the hidden layers act like the connections between neurons in the brain. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. Deep neural networks can thereby analyse various aspects of the input data at different stages to predict the output. In the training stage, the weights are adjusted until it reaches the optimal set of weights that predicts the output data with the highest accuracy.[8] Deep learning offers more flexibility than other ML approaches, but it needs larger training datasets (Table 1.I).[2, 4, 6, 7, 9]

Clinical applications

There are numerous ways in which ML can be applied to clinical tasks, covering all stages from raw-data acquisition to processing, analysis, diagnosis, prognosis, and treatment decisions.

Image processing

Precise measurements of anatomical structures such as arteries, veins and cardiac chambers are required for the investigation of many cardiovascular conditions, for example the structure of the left ventricle in hypertrophic cardiomyopathy[10] or the size of the aorta to detect an aneurysm.[11] Segmentation, the process of extracting an outline of an anatomical structure, is a time-consuming process when performed manually by humans which often shows considerable inter-rater variability.[12] Automation of segmentation with ML affords more rapid segmentation and precise quantification reducing inter-rater variability.[13] In addition to these benefits, images can automatically be checked for image quality, and novel diagnostic parameters such as cardiac shape metrics can be extracted from images, and anatomical maps showing diagnostic characteristics spatially can be created.[14]

Diagnosis

Supervised machine learning approaches are frequently applied to develop diagnostic tools. Supervised ML algorithms can be trained on images, for example a chest X-ray[15] or with a set of clinical parameters[16] or a combination of both. If model training relies upon input data labelled by a human expert, as can occur in supervised ML, the model will likely produce similar mistakes, or have similar biases, as the expert.[17] However, the diagnostic label may also be able to take into account information the clinician does not have access to at the time of diagnosis, for example longitudinal outcome data or a gold-standard test result. Theoretically this could improve diagnostic certainty of the ML model above that achievable by the clinician. For example, an invasive angiogram derived label concerning coronary artery lesions could be an appropriate label to train a ML coronary artery disease classifier based upon stress echocardiogram data.[18] Clustering approaches are also used for diagnostic purposes by grouping people together based upon data similarities which represent phenotypic characteristics. When a patient fits in a cluster with a high disease

Table 1.I. Overview of machine learning methods that are frequently used for cardiovascular clinical applications.

Method	Description	Primary reference
Supervised learning	**Training data in which the conditions or events are labelled is used to train the model to classify or predict these labels.**	**6**
Support vector machines	Training examples are mapped onto points in space to maximize the distance between the two categories. New examples are mapped into that same space and predicted to belong to the category they are closest to.	2
Decision trees (random forest)	Internal nodes split into branches representing decisions for which the training input most often leads to the correct label in the output. The predicted outcome is represented by the last branches of the tree that do not split any further (leaves or decisions).	6
Semi-supervised clustering	Class-uniform clusters that have high probability densities of training data labels	9
Unsupervised learning	**Uses data from unlabeled examples to identify groupings or outliers with no comparison to a predetermined or known outcome**	**4**
Principal component analysis	Reducing a large set of parameters into a low- dimensional representation by identifying the features that account for the most variation in a dataset.	2
Clustering	Training data is grouped by its similarity on a set of traits while minimizing the distance between data points.	6
Ensemble learning	**Uses a combination of models in three different ways to optimize performance beyond that obtainable with a single model.**	**6**
Reinforcement learning	**Processes dynamic, constantly changing data and, in responding to interactions with its environment, the model learns an optimized set of rules for achieving a goal or reward (or avoiding a penalty) by a process of trial and error.**	**6**
Deep learning	**Collective term for neural networks reminiscent of human brain synapses.**	**7**

prevalence, they are more likely to have this disease.[19] Similarly, clustering can be used to identify those with milder or more severe phenotypes of a condition.[20]

Prediction

ML models can not only be trained to detect whether a disease is currently present, but also whether a person is at risk of disease in the future. Prediction and risk stratification models have been widely used in clinical practice to aid in the therapeutical decision-making process. Such models require a training dataset that has longitudinal follow-up information available. Knowing risk of future disease can help stratify those with higher risk of cardiovascular disease to receive earlier intervention.[21] A ML method applied to predict cardiovascular events in >6800 asymptomatic individuals demonstrated a superior performance to the traditional CAC risk score.[22] When computational models are expanded to include information based on both current (diagnosis) and future (prognosis) heath status, treatments can be recommended based on a model-predicted projection of the pathways to restore health. This could be performed by using clustering analysis to predict a person's disease pathway by looking at the known pathways of people in the same cluster or by creating a "digital twin:" an *in-silico* representation of a person based on their medical data. The digital twin has a real-time connection between the person's medical data and the model. In combination with population representations, a digital twin would also allow for a simulation to predict which treatment option is the optimal choice.[23] With increasing pressures on healthcare services due to ageing populations in many developed countries, chronic disease and hospitalization prevention will become even more important.

Research

Computational models can identify the most important diagnostic data and reliably infer biomarkers that cannot be directly measured.[23] For example, in deep neural networks, connections with a higher weighting indicate that the information from the preceding nodes were most important for the task. Meanwhile in cluster analyses, the average values of a cluster on the principal components and the variables that are reflected in those principal components provide an indication of what original features were the most important. Knowing what phenotypical features are important for diagnosis and how these change over the clinical trajectory can be used to infer pathophysiological mechanisms. Based on phenotypical features, an unsupervised learning model has identified responders to cardiac resynchronization therapy in patients with heart failure.[24] Furthermore, models can reveal complex relationships between features that can lead to disease. Using ML methods for interpatient similarity analysis in cardiac structure and function a phenotypic map can be developed with specific characteristics and locations that differentiate cardiovascular disease stages and clinical outcomes.[19]

Structured reporting

To be able to use large amounts of clinical data from medical records that are historically or partly on paper, machine learning can help transcribe the paper records digitally.[4] Even when records are digital, they cannot always be readily used for computational modeling. In clinical practice, important information is often recorded in free text information. Summarizing this information into codified form for computer processing would be a

time-consuming process for healthcare professionals. Fortunately, this can now largely be automated using natural language processing.[1] A supervised deep neural network designed to predict in-hospital mortality in patients with cardiovascular disease using echocardiography report data and ICD-10 codes provided more accurate prediction compared to the existing prediction models (Figure 1.1).[25]

Data

Data sources

Machine learning models need datasets of sufficient quality, detail, and size to "learn" successfully. Big data is a term frequently coined to describe large amounts of collected data, often with high dimensionality. Medical health records are an important source of big data in medical research, as they contain both quantitative and qualitative data in numerical and textual formats from single encounters up to an entire lifetime. The benefits of this type of data are that it is already collected and therefore incurs less time, money, and effort to acquire than collecting new data. They also have the advantage of being real-world data, so can provide a more direct and relevant link between the research undertaken upon the

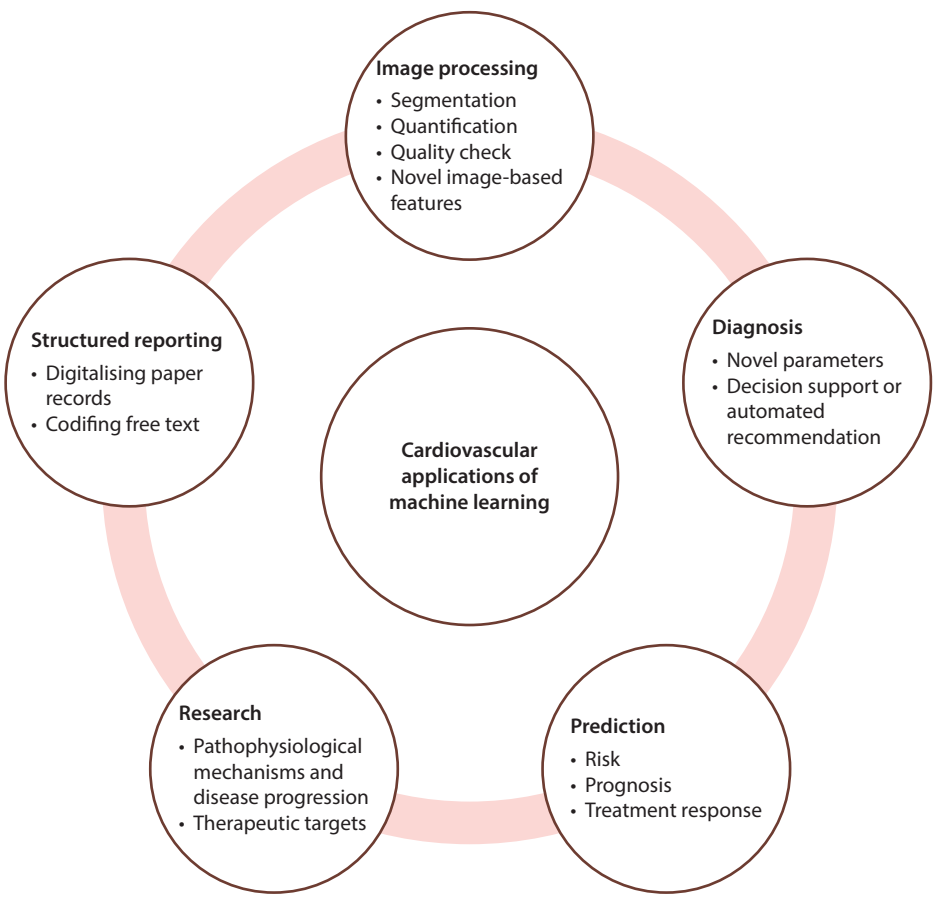

Figure 1.1. Categories of machine learning applications.

data and routine clinical practice. Since data is collected on every patient, there is less recruitment bias and therefore wider representation as compared to a research study with active recruitment. However, getting access to health records requires rigorous procedures to make sure that patients cannot be identified from their data. Since the data is not collected for research purposes, getting it up to standard for research analysis can be difficult and time-consuming. There are variations in the types of equipment used, how the data is collected and recorded between medical centers and even between clinicians within each center. Furthermore, the extent of data that is available is strictly limited to what the clinician deemed necessary for clinical evaluation at the time it was recorded. These limitations are largely overcome by biobanks, where data is collected by a research program with a high level of standardization and quality control. Large, population wide biobanks such as the United Kingdom (UK) Biobank aim for a broad representation of the population. However, participation is still limited by inclusion criteria (*e.g.*, age), cultural factors and participants' willingness and ability to commit to the time and travel involved in participation.[26] Biobanks offer a rich resource with which to train ML tools. Consortium studies are similar to biobanks in that they are a research program with predefined standardized data collection and quality control procedures, but often consist of research centers across multiple countries. Data from multiple studies can be collated into one database to create sufficient data with which to train ML tools. This approach requires fewer additional resources than collecting new data but has the limitation of variability between sites in terms of data acquisition, analysis, recording and formatting.

Missingness

In medical data, it is highly likely that some datapoints will be missing. This could be completely random (missing completely at random, MCAR) in which case a complete case analysis will not lead to biased results. If data is missing at random (MAR), the random probability of the missingness is dependent on the variables in the dataset and can be solved by re-weighting of the data. If data is missing not at random (MNAR), the missingness depends on the missing variable itself or on other missing and unobserved variables.[27] It is very important to understand the reasons behind missingness as it can reflect information and human biases. Ignoring missingness may lead to incorrect models that could potentially result in harmful predictions.[27]

Data sharing

ML is dependent on data and therefore data may need to be shared between health systems and those with the expertise to analyze the data such as computational scientists or biomedical engineers. Furthermore, data from multiple institutions may have to be shared to create datasets large enough to train and test ML algorithms. To protect the rights and identities of the people whose data is being shared, governments have set up legal data sharing frameworks. The European Union has General Data Protection Regulations, and the United States of America (USA) has the Health Insurance Portability and Accountability Act. Research study proposals must go through ethics approval processes, where designated organizations check whether all ethical requirements are met. This includes whether the

identity of participants is sufficiently protected. This is becoming increasingly difficult due to the processing of larger volumes of data, where not every record can be double checked for complete anonymization, and due to better algorithms that can identify persons with increasingly less information. On the other hand, there is a push for sharing data freely. Some academic journals and funders now require the data used in a publication to be open access to promote reproducibility testing and transparency.[1]

Data preparation

The quality of any ML application depends on the quality of the training data. As previously discussed, this data needs to be of sufficient size and quality at the collection stage, but to further prevent inconsistencies and errors, the data needs to be cleaned and further processed before it can be used. First, the data must be acquired using consistent methods of high standard and relevant parameters for the research question. Second, the data should be checked for missingness and data entry errors. When combining datasets from multiple studies or sites, extra care must be taken to ensure that there is no variability in data entry, for example in units (*e.g.*, L and mL), or formulae used (*e.g.*, different equations are available to calculate left ventricular mass).[28] If many datapoints are missing in a parameter, the decision can be made to remove this parameter from the dataset. Datapoints may also be imputed if a complete dataset is required for the ML modeling, as is often the case.

Modeling

Algorithm choice

The research aims and the available data drive the choice of a suitable ML algorithm(s). A supervised algorithm such as a support vector machine or decision tree would be appropriate for image segmentation, diagnosis and prediction applications where labelled data are available and the so called "ground-truth" is known. Unsupervised methods are better for identifying phenotypes or novel data patterns and may be the only choice if the data is unlabeled.[3] In other cases, a combination of algorithms may be required to train the model.

Performance metrics

Several metrics are commonly used in ML to assess model performance at the required task and to help drive the model training. Accuracy, the ratio between the correctly classified samples to the total number of samples, and error rate, the ratio of misclassified samples from both positive and negative classes to the total number of samples, are commonly used performance measures for supervised classification algorithms. Accuracy can be divided into sensitivity and specificity. Sensitivity, also called true positive rate, hit rate, or recall, reflects the proportion of the positive samples that were correctly classified. Specificity, otherwise known as true negative rate or inverse recall, represents the proportion of the negative samples that were correctly classified. The proportion of correctly classified positive samples to the total number of positive predicted samples is called the positive prediction value or precision. The proportion of correctly classified negative samples to the total number of negative predicted samples is called the negative predictive value, inverse