

# Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods

Hooman H. Rashidi, MD, FASCP<sup>1</sup>, Nam K. Tran, PhD, HCLD (ABB), FACB<sup>1</sup>, Elham Vali Betts, MD, FASCP<sup>1</sup> , Lydia P. Howell, MD, FASCP, FCAP<sup>1</sup>, and Ralph Green, MD, PhD, FASCP, FCAP, FRCPath<sup>1</sup>

## Abstract

Increased interest in the opportunities provided by artificial intelligence and machine learning has spawned a new field of health-care research. The new tools under development are targeting many aspects of medical practice, including changes to the practice of pathology and laboratory medicine. Optimal design in these powerful tools requires cross-disciplinary literacy, including basic knowledge and understanding of critical concepts that have traditionally been unfamiliar to pathologists and laboratorians. This review provides definitions and basic knowledge of machine learning categories (supervised, unsupervised, and reinforcement learning), introduces the underlying concept of the bias-variance trade-off as an important foundation in supervised machine learning, and discusses approaches to the supervised machine learning study design along with an overview and description of common supervised machine learning algorithms (linear regression, logistic regression, Naive Bayes, *k*-nearest neighbor, support vector machine, random forest, convolutional neural networks).

## Keywords

algorithms, artificial intelligence, convolutional neural network, deep learning, *k*-nearest neighbor, machine learning, random forest, supervised learning, supervised methods, support vector machine, unsupervised learning

Received April 28, 2019. Received revised July 15, 2019. Accepted for publication July 26, 2019.

## Introduction

Medical data are reported to be growing by as much as 48% each year.<sup>1</sup> This explosion of data and the associated challenges of its optimal use to improve patient care are driving development of a myriad of new tools that utilize artificial intelligence (AI) and machine learning (ML). Artificial intelligence is the capability for machines to imitate intelligent human behavior, while ML is an application of AI that allows computer systems to automatically learn from experience without explicit programming. Paraphrasing Arthur Samuel and others, ML models are constructed by a set of data points and trained through mathematical and statistical approaches that ultimately enable prediction of new previously unseen data without being explicitly programmed to do so.<sup>2,3</sup> Once residing

only in the realm of science fiction, advancements in computing power and accessibility has prompted a technological revolution involving AI and ML that is already impacting many domains of our everyday lives, including credit decisions, travel, personalized suggestions for movies, books, and other products as well as temperature control in our own homes.

<sup>1</sup> Department of Pathology and Laboratory Medicine, University of California Davis, School of Medicine, Davis, CA, USA

### Corresponding Authors:

Hooman H. Rashidi and Nam K. Tran, Department of Pathology and Laboratory Medicine, University of California Davis, 4400 V St, Sacramento, CA 95817, USA.

Emails: hrashidi@ucdavis.edu; nktran@ucdavis.edu



These tools are being increasingly incorporated into a broad range of clinical practice in many different medical disciplines and have become an area of intense investigation. Reflecting the growing role of AI/ML in medicine, the Food and Drug Administration recently issued a white paper<sup>4</sup> to safely guide AI development, which underscores the promise that AI and ML are believed to hold for improving medical practice and patient care.

The field of pathology and laboratory medicine is important to the development and ongoing improvement in many medical AI/ML tools and will likely play an even larger and more pivotal role as AI and ML applications expand across health-care settings. Perhaps as many as 70% of all medical decisions are based on laboratory tests.<sup>5</sup> Additionally, the bulk of data in the electronic medical record is from the clinical laboratory. Test results from pathology and the clinical laboratory frequently serve as the gold standard for clinical outcomes studies, clinical trials, and quality improvement. This massive amount of data requires enormous capacity for storage and sophisticated methods for handling and retrieval of information, necessitating the application of certain data science disciplines such as AI/ML.

Pathologists and laboratorians are therefore excited about the promise that AI/ML can bring to their ability to impact health care; however, even those interested in pursuing AI/ML as an area of clinical investigation or quality improvement are largely unfamiliar with the field and the processes involved with utilizing the tools it has to offer. The variable quality of medical and laboratory data available for use as well as the sheer diversity and complexity of ML algorithms creates a cornucopia of choices as well as challenges for investigators seeking to develop the best AI/ML predictive model. Once a quality data set has been established, an optimal ML model needs to be identified which means fully vetting the algorithms by building and testing multiple models for their appropriateness to the task at hand.

The most successful AI/ML models arise from multidisciplinary teams with expertise in ML, clinical medicine, pathology and laboratory medicine, biostatistics, and other relevant skillsets. Such a multidisciplinary team will be best equipped to address the following queries that are fundamental to successful project design:

- 1) Does the project address a need?
- 2) Is there sufficient data and is it the “right” kind of data that is both readily available and vetted by clinical experts in the field?
- 3) Which ML approach to use?
- 4) Are the optimized ML models applicable and generalizable when applied to a novel data set?

The purpose of this article is to facilitate cross-disciplinary literacy among pathologists, laboratorians, biomedical scientists, and individuals from other medical disciplines seeking to work in multidisciplinary teams to develop or facilitate the early adoption of AI/ML tools in health care. We define cross-

disciplinary literacy as having sufficient content knowledge (including strengths and limitations of availability tools and the concepts behind them) as well as a working understanding of the field’s unique vocabulary that interested individuals from other disciplines can understand written and spoken communications, think critically, and use this knowledge and skill in a meaningful way for their own discipline. To accomplish this goal, we describe the current landscape for AI/ML in pathology and laboratory medicine by defining the elements and numerous available options necessary to address the 4 queries essential to design of AI/ML tools in health care outlined above: defining the purpose, data curation and quality, choosing the most appropriate ML algorithm, and testing/validation. Table 1 is the glossary of commonly encountered ML terminology within the scope of this article which provides definitions and examples for each term.

## Current Landscape and Approach to Developing Machine Learning tools

### 1) Would the AI/ML tool address a real health-care need (defining the purpose)?

There is clearly a need to apply rational and systems-based data science principles for handling the ever-growing body of both qualitative and quantitative aspects of medical laboratory information and classification. Faced with the limitations of human processing of rapid, accurate, and precise retrieval of data in real time, the heuristic provided and amplified by ML offers an attractive approach to substantially improve the delivery of health care. Current health problems that are deemed suitable to ML include, but are not limited to, integrating multiple variables to mimic human clinical decision-making skills (eg, multiparameter disease diagnosis), automation of testing and treatment algorithms (eg, reflex testing) and workflows, pattern recognition using imaging data (eg, radiology, histology slides, and vital sign waveforms), and/or test utilization trends. However, although one could use AI/ML, it may not always be necessary to apply such tools for every situation since simple statistical approaches may sometimes suffice.

### 2) Is there sufficient data and is it the “right” kind of data that is both readily available and its quality verified?

The familiar concept of “garbage-in/garbage-out” highlights the critical importance of having high-quality data for AI/ML applications, since incomplete and/or erroneous values may inappropriately train an algorithm in the wrong direction. Likewise, highly controlled data may not represent real-world conditions. “Quality data” for AI/ML training applications must include accurate, precise, complete, and generalizable information.<sup>6</sup> Laboratory data are often assumed to be sufficiently accurate and precise by both health-care providers and researchers. Unfortunately, it is a truism that not all laboratory tests are created equal, and poor analytical bias and imprecision

**Table 1.** Common Machine Learning Terms.

Term	Definition	Example(s)									
Bias-variance trade off	This refers to finding the right balance between bias and variance in a machine learning (ML) model, with the ultimate goal of finding the most generalizable model. Notably, increased bias usually leads to an underfitted model while increased variance may lead to overfitting. Finding the happy balance between the bias and variance is the key to finding the most generalizable model	See Figure 2									
Boosted tree	An ensemble method that uses weak predictors (eg, decision trees) that can ultimately be boosted and lead to a better performing model (ie, the boosted tree).	Also, Gradient boosting machine (GBM)									
Bootstrapping	To randomly pull samples from the original data set for creating a new data set of the same size. Note: "Bagging" is a term that is related to this which stands for Bootstrapping aggregation. This ensemble approach uses the aggregated data to make a decision by initially building multiple models on a subset of data whose predictions are ultimately then combined to make the final prediction	Regularly used in certain ensemble decision tree algorithms such as random forest									
Categorical data/features	These include features that have discrete values and are often binary. Although not infrequently, more than 2 categories have also been used	Patients with diabetes and patients without diabetes									
Class	Refers to the labeled target values	For a binary classification of cancer diagnosis, the classes could be cancer versus no cancer designated as the targets									
Classification modeling	A supervised machine learning approach that builds models that are able to distinguish 2 or more discrete classes	Cancer versus no cancer									
Classification threshold	This is usually the probability/value threshold that allows the model to be able to separate the 2 classes.	If one has used the default probability value of 0.5 in our logistic regression-based model to separate the positive cases of cancer from the negative cases of cancer, the positive cases are those that are $\geq 0.5$ and the negative cases are those that are $< 0.5$ .									
Clustering	Refers to grouping related data points. This is commonly seen in certain unsupervised machine learning methods.	k-means algorithm (see K entry terms in the glossary below)									
Confusion matrix	Summarizes the model's predictions. Specifically denoting the true positive, true negative, false-positive, and false-negative predictions. These values will then make it possible to calculate the various performance measures of our model (eg, accuracy, sensitivity, specificity, etc)	<table border="1"> <thead> <tr> <th></th> <th>Predicted Cancer Cases</th> <th>Predicted Negative Cases</th> </tr> </thead> <tbody> <tr> <th>Actual cancer cases</th> <td>True positive</td> <td>False Negative</td> </tr> <tr> <th>Actual negative cases</th> <td>False positive</td> <td>True Negative</td> </tr> </tbody> </table>		Predicted Cancer Cases	Predicted Negative Cases	Actual cancer cases	True positive	False Negative	Actual negative cases	False positive	True Negative
	Predicted Cancer Cases	Predicted Negative Cases									
Actual cancer cases	True positive	False Negative									
Actual negative cases	False positive	True Negative									
Convolutional Neural network (CNN)	These are usually deep neural networks with at least 1 convolutional layer and that are well suited for certain complex tasks such as image recognition/classification.	See Figure 3									
Cross Industry Standard Process for Data Mining (CRISP-DM)	A systematic approach to supervised machine learning process that includes the following steps: Data collection and processing, followed by model building and validation steps and ultimately the model deployment step	See Figure 4									
Cross validation	A way to statistically estimate the model's generalizability by withholding 1 or more internal test sets that can then be tested against the trained model(s).	k-fold cross validation (see K entry terms in the glossary below)									
Data types	Data within ML platforms are best categorized into the following 4 types: numerical (exact numbers), text (which will need to be converted to numbers), categorical (represents characteristics such as normal tissue vs cancer, etc), and time series (sequence of numbers collected over time intervals)	An example of categorical is normal tissue versus cancer cases. Note: 7 data type grouping has also been proposed which include: useless, nominal, binary, ordinal, count, time and interval data.									
Decision boundary	This refers to the boundary between the classes that is learned by the machine learning model. In the example, the blue dashed lines represent the best decision boundaries that separate the 3 classes (black group, gray group, and the red group).	See Figure 3									

Table 1. (continued)

Term	Definition	Example(s)
Decision tree	These use a flowchart structure that typically contains a root, internal nodes, branches, and leaves. The internal node is where the attribute in question (eg, creatinine > 1 or creatinine < 1) is tested while the branch is where the outcome of this tested question is then delegated. The leaves are where the final class label is assigned which, in short represents the final decision after the results of all the attributes have been incorporated. The end result of the decision tree is a set of rules that governs the path from the root to the leaves	Simple Decision tree (not often used) Boosted Tree (Gradient Boosting Machine) Random forest (ensemble of decision trees)
Deep Neural Network (DNN)	Refers to a neural network with multiple hidden layers and a large number of nodal connections	Common deep neural networks that are currently used within the image analysis field include AlexNet, Inception which is also known as GoogleLeNet (eg, Inception-v3) and ResNet (eg, ResNet50). These are commonly employed in various transfer learning projects
Discrete	Qualitative targets or features within a classification schema in supervised learning.	Examples of discrete targets may include cancer versus normal tissue or acute kidney injury (AKI) versus No-AKI in a classification ML model. This is in contrast to quantitative targets that can be used within supervised learning models and that can then also be used to predict a numerical outcome as in a linear regression model
Feature	Refers to the input variables that are used to map to the target in a model	For example, in an acute kidney injury (AKI) model certain features such as, creatinine, urine output, and NGAL may be used to map to its Target categories (AKI vs No-AKI), which would ultimately make it possible to build a model that can predict AKI
Feature engineering	This the process that allows for the selection of certain key features or the transformation/ conversion of certain features within a data set that will ultimately lead to a better prediction model	A data set with a large number of features can be refined by certain methods (eg, PCA or K-Best, etc) to find the most relevant features (within all features) based on various statistical methods. This can then be used within the algorithm to build a new model. For example, a data set with 20 features or more can be refined through PCA or K-Best to a new data set with 10 or less features
Generalizability	The ability of a model to accurately predict on new previously unseen test sets (secondary or tertiary test sets) that were completely outside of the training data set.	In machine learning, the validation accuracy typically refers to the model's accuracy based on the primary test set that was created with the data set that was used to train the model within its train-test split phase. In contrast, the generalization accuracy is a new test set that is used to test the final model and its capability of predicting using previously unseen data.
Generalization test set	The secondary or tertiary test set that is unknown to the initial train-test split data set and used to assess the model's generalizability.	A cancer-predicting model trained with images from one institution can be tested for its generalizability using another institution's images as a secondary test set.
Input variables	These usually refer to the features from the training data.	In a cancer predicting model trained with cancer and no-cancer images, the image features within cancer are used to map to the cancer-labeled group while the image features of the benign group are mapped to the no-cancer group.

(continued)

**Table 1.** (continued)

Term	Definition	Example(s)
K-fold cross validation (CV)	A type of cross validation in which the train-test data set is split k times.	If the k is 10 in the k-fold CV, the data are split into 10 train-test splits to assure proper sampling of the training set and more importantly of the testing data.
K-means	An unsupervised method that utilizes discrete or continuous data as its input parameter for identifying input regularities (ie, clusters).	Clustering a medical data set to characterize subpopulations of a disease based on various known input clinical parameters. See Figure 3
K-nearest neighbor (KNN)	A nonparametric clustering algorithm used for data classification and regression. Classification is based on the number of k neighbors, where k is equal to the square root of the number of instances, and its distance (eg, Euclidean) from a predefined point.	See Figure 3 (SVM example)
Kernel trick	Allows the data to be transformed into another dimension (eg, z-plane) which ultimately enhances the dividing margin between the classes of interest.	
Leave One Out cross validation	The extreme version of the k-fold CV approach in which k will equal to n (eg, total number patients being studied). Instead of train-test, splitting them into k-folds of 5 or 10 (as in k-fold CV studies). As the name implies, by leaving one out each time for each cross-validation step, k will then equal to the total number of individual data entries n (eg, n of 100 could be the individual sets of data in 100 patients).	For example, n of 100 could be the individual sets of data in 100 patients. Hence, the k-fold CV in this case will split the data into 100 train-test splits which assure full sampling
Linear regression	This algorithm allow us to find the target variable (usually a numerical value) by finding the best-fitted straight line which is also known as the "least squares regression line" (the best dotted line with the lowest error sum) between the independent variables (the cause or features) and dependent variables (the effect or target). The ultimate goal of this technique is to fit a straight line to the data set in question.	See Figure 3
Logistic regression	The term regression is somewhat of a misnomer since in general this is a classification method that uses a logistic function for predicting a dichotomous dependent variable (target).	See Figure 3
Loss	It measures how poorly the model is performing when comparing its predictions to the target labels. Hence, typically the lower the loss, the better the model.	For example, measuring "log loss" in logistic regression models and "mean squared error" (MSE) in linear regression models.
Machine learning (ML)	Machine-based intelligence (in contrast to natural human intelligence). Also interchangeably used with the term artificial intelligence (AI). Paraphrasing Arthur Samuel and others, ML models are built by a set of data points trained through mathematical and statistical approaches that ultimately enable prediction of new previously unseen data without being explicitly programmed.	Supervised ML: Classification and Regression models Unsupervised ML: Clustering and Dimensionality Reduction (eg principal component analysis, see below) Reinforcement Learning
Model	This usually refers to the end result of the machine learning algorithm's training phase in which the variables are ultimately mapped to the desired target.	A deep neural network model that is trained to predict cancer from benign tissue or a k-NN model that is used to predict acute kidney injury (AKI) from no-AKI See Figure 3
Naive Bayes	A classifier that uses a probabilistic approach based on the Bayes theorem. This approach assumes the naïve notion that the features being evaluated are independent of each other.	Apple's Siri; Amazon's Alexa
Natural Language Processing (NLP)	An ML process that enables computers to learn and ultimately analyze human (natural) language data. This may include various aspects of speech, syntax, semantics, and discourse (oral and written).	
Normalization	A standardization or scaling procedure that is often used in the processing phase of the data in preparation for machine learning. Certain methods (eg, distance-based algorithms) within machine learning (eg, k-NN) are very sensitive to unscaled data. Hence, in the absence of scaling (ie, normalization) they may not perform as expected. Usually, normalizing the data for such algorithms improves their overall performance	For example, Standard scaler (scales the data so that the mean of the distribution is set to 0 with a standard deviation of 1.)

(continued)

6 **Table 1.** (continued)

Term	Definition	Example(s)
Overfitting	This gives rise to the model appearing as a good predictor on the training data while underperforming on future new and previously unseen data (ie, not generalizable). This is due to its a low bias and high variance in which the model may now adapt too strongly to the data which could have included noise.	See Figure 2
Parametric algorithm	The set of parameters in a parametric algorithm is fixed which confines the function to a known form. In general, the assumption within parametric algorithms is that the function is linear or assumes a normal distribution while nonparametric methods do not make such assumptions	The most commonly encountered parametric algorithms include linear regression, logistic regression, and naive Bayes, while some of the most common nonparametric algorithms include k-NN, SVM, CNN, and decision trees including RF
Prediction	Refers to the model's output based on some initial input variables	A deep neural network trained to identify cancer cases is tested against an unknown histologic image which predicts it as cancer with a certain probability.
Principal Component Analysis (PCA)	A statistical approach that can lower dimensional representations. This technique can highlight the contribution of various features within a data set through its principal components which could ultimately allow a reduction in the number of features that are required to build the model.	Within a data set that contains, say, 8 features, 92% of the explained variance within the data may have been found to be contained within the first 2 principal components (PC1 and PC2) which were subsequently shown to be mainly due to 4 of the 8 features. Then, those 4 features may be selected to build an ML model if one is looking to train a model with a smaller number of features
Random Forest (RF)	Uses a network of decision trees for ensemble learning. Using bootstrapping, this method generates randomly generated data sets that can then be used to train the data for building an ensemble of decision trees. Ultimately, each decision tree will determine an outcome, and a majority "vote" approach is used to classify the data. Appropriately, this is called random forest since a large number of randomly generated decision trees are used to construct the final model.	See Figure 3
Regularization	A process that helps to minimize the overfitting of an ML model by reducing the effect of noise.	This usually refers to adding some information to the process that ultimately reduces overfitting and makes the model a better predicting tool.
Reinforcement learning	These platforms may share features of both a supervised and an unsupervised process and usually function through a policy-based platform.	IBM's Watson and Google's Go that were able to beat champion chess and Go players, respectively.
Scikit learn	A popular machine learning library that enables users to build and assess various ML models through a variety of supervised and unsupervised algorithms. Other commonly used ML libraries (besides Scikit learn) that are especially useful for image classification model building include Turi create and Tensor flow.	Building a classification model through its support vector machine or logistic regression algorithm.
Supervised learning	These platforms employ "labeled" training data sets (labeled/supervised by subject experts) to yield a qualitative or quantitative output. The 2 major categories of supervised learning are classification and regression which lead to discrete/qualitative and continuous/quantitative targets, respectively.	Classification models; regression models
Support Vector Machine (SVM)	Classifies data by defining a hyperplane that best differentiates 2 groups. This differentiation is maximized by increasing the margin (the distance) on either side of this hyperplane. In the end, the hyperplane-bounded region with the largest possible margin is used for the analysis. One of the key highlights of the SVM method is its ability to find nonlinear relationships through the use of a kernel function (kernel trick).	See Figure 3
Target	Within supervised ML, the target is also sometimes referred to as the label which is comprised of the results or classes that one seeks to find.	In a supervised ML model for cancer versus normal tissue, the "cancer" and "normal tissue" are the labels (targets).

(continued)

**Table 1.** (continued)

Term	Definition	Example(s)
Train-Test Split	A common approach employed in supervised machine learning in which a subset of the initial data is used to train the model and a subset that is set aside is used to test its initial validation.	A 80–20 train-test split is one in which 80% of the initial data set is assigned to the training phase and 20% kept behind and used to test the model's performance.
Transfer learning	In this method unrelated images (eg, cancer vs benign histology) are retrained into a preestablished convolutional neural network (eg, ResNet-50) that is usually devoid of such data.	This approach can be used to build accurate ML models that can distinguish histologic variants of disease by retraining a preestablished neural network such as ResNet50 or Inception-v3.
Underfitting	An underfitted model has a higher bias and lower variance. In this situation, important potential interrelationships between the data features may be ignored.	See Figure 2
Unsupervised learning	These involve agnostic aggregation of unlabeled data sets yielding groups or clusters of entities with shared similarities that may be unknown to the user prior to the analysis step.	Clustering; Dimensionality Reduction (eg, PCA)
Validation testing	This usually refers to the initial validation testing phase in which the test set (eg, 20% of the initial data set) that was set aside from the train-test split is used to assess the model's initial performance. This does not always correlate with the generalizability of the model. Hence, testing the model with secondary test sets is usually recommended.	In a model in which 20% of the initial data set was kept behind and used to test the model's performance, the model was found to be 92% accurate (its validation accuracy). However, using a secondary test set the model was found to be 81% accurate (the model's likely generalizability).

Abbreviations: AKI, acute kidney injury; CV, cross-validation; k-NN, k-nearest neighbor; ML, machine learning; NGAL, neutrophil gelatinase-associated lipocalin; PCA, principal component analysis; RF, ensemble decision tree algorithm; random forest; SVM, support vector machine.

degrade the performance of AI/ML algorithms. Additionally, both providers and researchers are often not aware that test methods may lack standardization. For instance, a cardiac troponin I assay from one manufacturer may not be the same when compared to another due to differences in epitopes for antibody-based capture and detection.<sup>7</sup> The concept of imprecision reported as coefficient of variation is also poorly understood by most bedside providers with many assuming any change in numerical values reflecting a true biological change without taking into account sources of variability.

Data completeness and generalizability are other important considerations when developing and training AI/ML algorithms.<sup>8</sup> Unfortunately, despite the convenience of collecting real-world information from electronic health records, the retrieved medical data are often incomplete. This is attributed to the several inconsistencies in test ordering and resulting. Ordered laboratory tests may be cancelled due to patients not showing up for a visit, or samples were found to be not acceptable upon receipt by the laboratory. Incomplete data create significant challenges for AI/ML developers, where the predictive power of algorithms may be severely diminished. The limitation of real-world evidence has thus prompted investigators to gravitate toward more complete and rigorous data derived from clinical trials. However, caution is advised when using data that are “too complete” or “too controlled,” since it may not represent the real-world population and contribute to overfitting, discussed later in this article.<sup>9</sup>

Ultimately, the best and most balanced approach is to pilot AI/ML algorithms using more controlled data during the initial stages and later refining these algorithms using real-world data to confirm generalizability.

### 3) Which ML approach to use?

Choosing the right ML approach for a given task requires a basic understanding of the general categories of ML algorithms as well as a basic understanding of these algorithms’ inner workings, strengths, and limitations. These are outlined below.

## Machine Learning General Categories: The Big Picture

Within the various ML platforms, there are a multitude of algorithms to choose from.<sup>10,11</sup> The choice of an algorithm depends on a variety of factors that include, but are not limited to, data type/learning approach (supervised or unsupervised learning), the need for  $k$  (clustering), the importance of accuracy in the chosen model, the need for speed in data analysis, the data analyzed, the size of the data set, the need for hierarchical output, and the need for categorical variables (Figure 1). Machine learning methods and algorithms belong to one of the following 3 categories: (1) supervised learning, including classification and regression approaches; (2) unsupervised learning<sup>12</sup>; and (3) reinforcement learning (Figure 1).

A supervised ML algorithm makes use of the training data to learn a function ( $f$ ) by mapping certain input variables/features ( $X$ ) from the training data into some output/target ( $Y$ ). In

general, supervised ML platforms employ “labeled” training data sets to yield a qualitative or quantitative output. The labeled nature of the data evaluated in the training phase is a key feature of this method, since it allows the ML model to ultimately emulate the expert’s input data. As a result, the ML model can distinguish an unknown input based on its prior training parameters. In the “classification” approach of supervised learning, the labeled data/variables (which can be numbers, text, or unstructured data such as images) yield a discrete (qualitative) “class” output. An example of a classification approach is the breast cancer histology image identification model in which a supervised ML platform is used to yield a qualitative answer/identification based on labeled histologic image training data sets that are then used to predict future unknown histologic images. In contrast, the “regression” approach of supervised learning involves the cumulative acquisition of data variables to yield a continuous (quantitative) numerical output (Figure 1). Notably, most reproducible supervised studies follow the Cross Industry Standard Process for Data Mining or some modification thereof.<sup>13,14</sup>

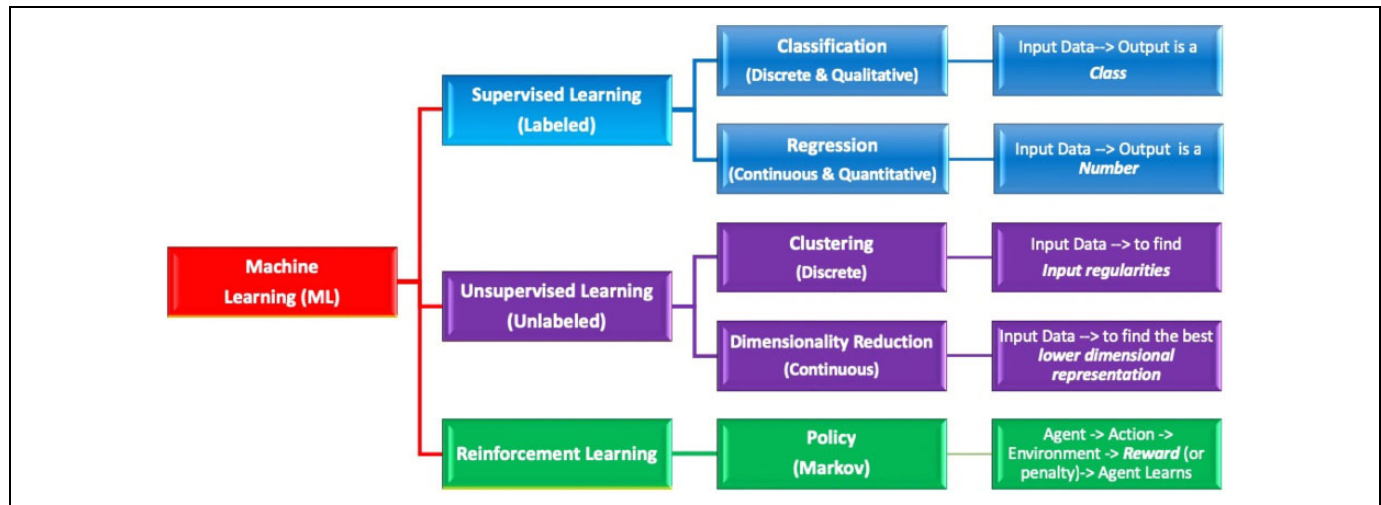
Unsupervised ML methods involve agnostic aggregation of unlabeled data sets yielding groups or clusters of entities with shared similarities that may be unknown to the user prior to the analysis step. These are also sometimes referred to as clustering algorithms. Some of the most common methods employed in this approach include  $k$ -means clustering, anomaly detection, or certain statistical methods such as principal component analysis.<sup>15-17</sup> These approaches usually utilize discrete or continuous data as their input parameter for identifying input regularities (eg,  $k$ -means clustering) or for lowering dimensional representations (eg, principal component analysis). An example is the use of ML to cluster unorganized/unlabeled laboratory data with no obvious commonalities into something new and meaningful for the user. Notably, the outcomes of various unsupervised ML methods (eg, results of a principal component analysis) can often also complement and thus enhance the performance of certain supervised learning ML methodologies.

Reinforcement learning platforms may share features of both a supervised and an unsupervised process and usually function through a policy-based platform. An example of reinforcement learning is International Business Machine (IBM)’s Deep Blue (Armonk, New York) and Google’s Go (Alphabet, Mountain View, California) that were able to beat champion chess<sup>18</sup> and Go players,<sup>19</sup> respectively. However, currently reinforcement learning approaches are rarely employed in pathology. This may change in the future.

## “Supervised” Machine Learning Algorithms (General Overview)

As noted earlier, in medicine and in pathology in particular, ML models employed are chiefly based on supervised approaches. Based on the amount of data and data type (eg, image vs numerical values vs text), the type of algorithm employed could drastically alter the ML model’s predictive





**Figure 1.** Overview diagram of machine learning algorithms. Machine learning is a subset of artificial intelligence. This figure illustrates the hierarchy of different machine learning algorithms including supervised versus unsupervised versus reinforcement learning techniques. The 2 major categories of supervised learning are classification and regression which lead to discrete/qualitative and continuous/quantitative targets, respectively.

capabilities. In the sections below, the various supervised algorithms within ML are discussed with an emphasis on the classification approach within the supervised learning category. The advantages and limitations of each are also provided since these provide insight into the approach to such studies.

The type of the input data can alter the approach to the analysis step and the type of algorithm that needs to be employed. Although similar algorithms may be applied to different data types, commonly used data types in the health sciences include image and text which have made use of visual recognition platforms and natural language processing frameworks, respectively. In both of these settings, deep learning neural network algorithms are now commonly employed. Deep neural networks have become the gold standard for image classification. However, neural networks are not the only algorithms within ML and may not always be the most suitable method when using nonimage data (eg, numerical laboratory data).

Commonly used supervised learning algorithms encompass both convolutional neural networks (CNNs; eg, deep learning) and various non-neural network algorithms (Table 2). Some of the most common non-neural network algorithms employed include linear regression, logistic regression, naive Bayes, decision tree,  $k$ -nearest neighbor ( $k$ -NN), support vector machine (SVM), and the ensemble decision tree algorithm random forest (RF).

In supervised classification platforms, if accuracy is not the ultimate goal, algorithms such as logistic regression or naive Bayes may suffice. However, if accuracy is the primary objective in these classification tasks, then the algorithms of choice currently include kernel SVM,  $k$ -NN, boosted tree, RF, and CNNs (especially deep learning). As noted earlier, the method of choice for most image classification tasks is now deep neural networks, which are typically CNNs with a large number of

artificial neural connections within their hidden layers. More importantly, these CNNs are not routinely built from scratch but rather retrained based on a transfer learning approach from preestablished neural networks. In transfer learning, unrelated images (eg, cancer vs benign histology) are retrained into a preestablished CNN (eg, ResNet-50) that is usually devoid of such data.<sup>20</sup> This approach is currently very popular and can be used to build accurate ML models that can distinguish histologic variants of disease in a relatively rapid pace.

On the other hand, in supervised regression (nonclassification) platforms, if accuracy is not the ultimate goal, algorithms such as the linear regression and decision tree may suffice. In contrast, if accuracy is the primary objective, then the algorithms of choice currently include RF and CNNs (Table 2).

### Bias-Variance Trade-Off in “Supervised” Machine Learning: A Fundamental Concept

The concept of bias and variance and their relationship with each other is fundamental to the true performance of supervised ML models. To identify the most optimized supervised ML model, the trade-off between bias and variance must be addressed. Briefly, bias gives the algorithm its rigidity while variance gives it its flexibility.<sup>21-23</sup> A high bias causes underfitting; simply stated, this means missing real relationships between the features of the data set and the target. In contrast, a high variance causes overfitting which may be thought of as introducing false relationships due to increased noise between the data set features and the target.<sup>24</sup> Thus, overfitting gives rise to the model appearing as a good predictor on the training data while underperforming on future new and previously unseen data (ie, not generalizable). In the end, the ultimate goal of any ML algorithm is to find the right balance between bias and variance (bias-variance trade-off). This balance is key in

**Table 2.** Comparison of Most Common Supervised Learning Algorithms.

Algorithm	General Accuracy of the Models Built	Used for Classification or Regression Tasks	Training Time	Algorithm Is Relatively Transparent	Able to Deal With Noise (to Tune out Irrelevant Features)	Need for Scaling/Normalization of Data	Highlights	Limitations
Linear regression	Low-intermediate	Regression	Rapid	Yes	No	Yes	Well studied and well known	Less predictive on closely correlated variables Sensitive to background noise May be limited by high number of features The assumption that the features are independent may be false Risk of overfitting
Logistic Regression	Low-Intermediate	Classification	Rapid	Yes	No	Yes	Simple Well studied and well known	
Naïve Bayes	Low-Intermediate	Classification	Rapid	Yes	Yes	No	Simple Very transparent Very Rapid	
K-Nearest Neighbor	Intermediate	Both	Rapid	Yes	No	Yes	Able to find nonlinear relationship	Risk of overfitting
Decision Tree (eg, Boosted Tree)	Intermediate-High	Both	Rapid	No	No	No	No real training process is required (grouped based on data distances)	The type of distance metric used may alter the performance Risk of overfitting
Random Forest	High	Both	Intermediate	No	Yes	No	Able to find nonlinear relationships Able to find nonlinear relationships	Risk of overfitting
Support Vector Machine	High	Classification*	Rapid	Yes	Yes	Yes	Able to find nonlinear relationships	Risk of overfitting
Convolutional Neural Network	High	Both	Slower	No	Yes	Yes	Method of choice for many image analysis studies Able to find non-linear relationships	Black box algorithm Risk of overfitting Requires higher computational power and time

\*Support Vector Regression (SVR), not discussed here, is the counterpart to SVM and used for regression studies.

finding the most generalizable model (Figure 2). Within many supervised ML approaches, with the appropriate test sets, this balance can be intrinsically automated and sometimes incorporated into the platform to ultimately identify the most suitable model. Being aware of such limitations and knowing how to appropriately approach these platforms for building the most suitable model is key to good ML practice.

## Supervised Machine Learning Algorithms: (Common Algorithms and Their Inner Workings)

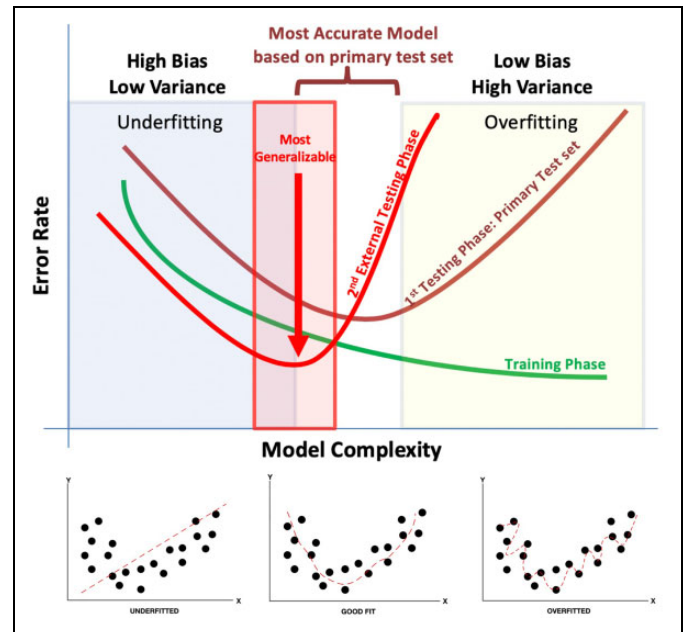
In addition to the abovementioned categorizations, the algorithms can be further divided into either parametric or nonparametric groups.<sup>25</sup> The set of parameters in a parametric algorithm is fixed which confines the function to a known form. In nonparametric methods, the algorithm does not make any assumptions about the function to which it will map its variables. In general, the assumption within parametric algorithms is that the function is linear or assumes a normal distribution, while nonparametric methods do not make such assumptions. The most commonly encountered parametric algorithms include linear regression, logistic regression, and naive Bayes, while some of the most common nonparametric algorithms include  $k$ -NN, SVM, CNN, and decision trees including RF (Figure 3). A small description of the inner workings of these algorithms along with highlights and limitations for each are discussed below and also included in Table 2.

### Linear Regression Algorithm

One of the oldest and simplest parametric statistical approaches is least squares linear regression. This technique has been regularly used for various correlational studies.<sup>26</sup> Linear regression models allow us to find the target variable (usually a numerical value) by finding the best-fitted straight line that is also known as the “least squares regression line” (the best dotted line with the lowest error sum) between the independent variables (the cause or features) and the dependent variables (the effect or target). The ultimate goal of this technique is to fit a straight line to the data set in question (Figure 3). The advantage of such an approach is its simplicity and transparency for finding linear relationship that can ultimately be very efficient (rapid). However, its major limitation is not being generally useful when relationships between the independent variable (the cause or features) and the dependent variables (the effect or target) are nonlinear.

### Logistic Regression Algorithm

The term regression is somewhat of a misnomer since in general this is a classification method that uses a logistic function for predicting a dichotomous dependent variable (target). A variation of this method (multinomial logistic regression) can also be used to classify more than 2 targets.<sup>27,28</sup> In the binary approach, the function yields a value of 0 or 1 which represents

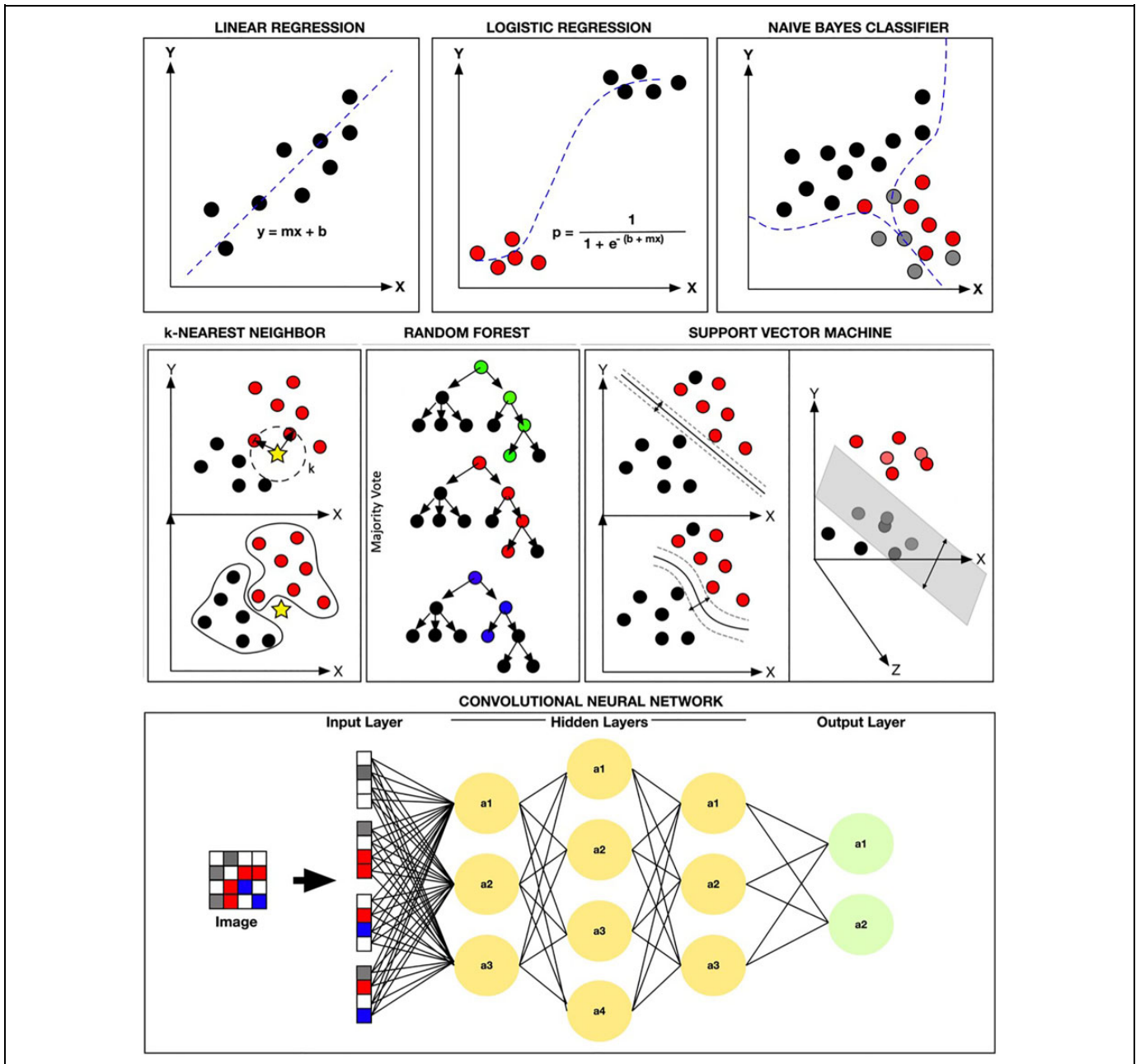


**Figure 2.** Bias-variance trade-off in machine learning. This figure illustrates the trade-off between bias and variance. Training data (green line) often do not completely represent results from the testing phase. Underfitting data are less variable but exhibit a high error rate and high bias (blue box). In contrast, overfitting data result in low bias and high variance (yellow box). The ideal zone lies between over-versus underfitting of data and may not be optimal until several attempts at testing have been made (red line).

the negative (0) and the positive (1) case (Figure 3). This may be accomplished by calculating an odds ratio probability for assigning a value as positive (1) or negative (0) based on the relationships between the independent input variables (features) and the dependent variables (target). This algorithm is relatively popular and has been regularly used in both industry<sup>29</sup> and medicine.<sup>30</sup> The use of a logistic regression method may become limiting if there are large number of features/variables present or if the variables are highly correlated. Additionally, this approach assumes that the relationship between the independent variables (features) and the dependent variables (target) are uniform which may limit the model’s performance.<sup>31,32</sup>

### Naive Bayes

Naive Bayes classifiers use a probabilistic approach that is based on the Bayes theorem. This approach is a subset of the Bayesian logic that assumes the naive notion that the features being evaluated are independent of each other.<sup>33-35</sup> Although this basic assumption may seem to be a disadvantage of this method, in reality, naive Bayes classifiers can sometimes yield reasonable results,<sup>34</sup> especially for simple tasks. However, their performance has been shown to be inferior to some of the other well-established algorithms such as boosted trees and RF.<sup>10</sup>



**Figure 3.** Comparison of popular supervised learning methodologies. This figure illustrates a variety of popular supervised machine learning (ML) methodologies. In the top row, linear regression, logistic regression, and Naïve Bayes Classifier (via TensorFlow) are shown. In the second row,  $k$ -nearest neighbor ( $k$ -NN), the ensemble decision tree algorithm random forest (RF), and support vector machine (SVM) are compared. Finally, the bottom row illustrates a convoluted neural network evaluating an image. Each image pixel is evaluated (input layer). The network contains several “hidden layers” (yellow circles) which is then processed and sent to the output layer (green circles).

### Decision Tree and Boosted Tree (Gradient Boosting Machine)

A decision tree uses a flowchart structure that typically contains a root, internal nodes, branches, and leaves. The internal node is where the attribute in question (eg, creatinine  $>1$  or creatinine  $<1$ ) is tested, while the branch is where the outcome of this tested question is then delegated. The leaves are where the final class label is assigned which, in short, represents the final

decision after it has incorporated the results of all the attributes.<sup>36-39</sup> The end result of the decision tree is a set of rules that governs the path from the root to the leaves. Simple decision trees are not commonly used in ML. However, variations such as the Gradient boosting machine is used for both classification and regression tasks.<sup>40,41</sup> Gradient boosting machine is an ensemble method that uses weak predictors (eg, decision trees) that can ultimately be boosted and lead to a better performing model (ie, the boosted tree). This method can

sometimes yield very reasonable models, especially with unbalanced data sets. However, their limited number of tuning parameters may make them more prone to overfitting compared to RF that contains a larger number of parameters for tuning and finding the optimized model.

### *k*-Nearest Neighbor

*k*-nearest neighbor (*k*-NN) is a nonparametric clustering algorithm used for data classification and regression. Classification is based on the number of *k* neighbors, where *k* is equal to the square root of the number of instances and its distance (eg, Euclidean) from a predefined point<sup>42-46</sup> (Figure 3). An ideal set of *k*-values may be identified that best predicts a desired outcome.<sup>11</sup> The use of *k*-NNs are relatively intuitive and as a nonparametric approach makes no assumptions when data points are assigned to a respective class. Additionally, they can be applied to both classification and regression tasks. However, they work best with a smaller number of input variables, they require feature scaling/normalization (since they are distance based), and they are sensitive to outliers within the data set.

### Support Vector Machine

Support vector machine classifies data by defining a hyperplane that best differentiates 2 groups. This differentiation is maximized by increasing the margin (the distance) on either side of this hyperplane. In the end, the hyperplane-bounded region with the largest possible margin is used for analysis.<sup>47</sup> One of the key highlights of the SVM method is its ability to find nonlinear relationships through the use of a kernel function (kernel trick). In short, the kernel trick allows the data to be transformed into another dimension which ultimately enhances the dividing margin between the classes of interest<sup>48</sup> (Figure 3). The limitation of this method is its tendency for overfitting.

### Random Forest

Random forest uses a network of decision trees for ensemble learning. Bootstrap technique is commonly employed in this method to generate the randomly generated data sets that can then be used to train the data for the ensemble of decision trees.<sup>49</sup> Ultimately, each decision tree will determine an outcome, and a majority “vote” approach is used to classify the data (Figure 3). Appropriately, this is called RF, since a large number of randomly generated decision trees are used to construct the final model.<sup>50-54</sup> This random sampling generally enhances the generalizability of this ML process by minimizing the overfitting phenomena. The number of trees and various other internal parameters within this process may hinder its performance. Additionally, the number of variables evaluated may be more time-consuming using this approach compared with the other nonparametric (eg, SVM and *k*-NN) and parametric methods (eg, logistic regression).

### Convolutional Neural Network

Neural networks attempt to emulate the neuron and for that matter the human brain. The artificial neuron within neural networks uses certain input features/variables to find and assign appropriate mathematical weights that are ultimately able to predict some output target (Figure 3). A deep neural network usually refers to a neural network with a large number of nodal connections within its hidden layer, and the CNNs are typically the deep neural networks that are most suitable for more complex data analyses such as imagery. As noted, in most CNN studies, a transfer learning approach is employed which allows the training data of interest to be incorporated into a retrained preestablished CNN.<sup>20,55</sup> The CNNs with the transfer learning approach are the method of choice for most image analysis studies. However, they are also prone to overfitting similar to the aforementioned algorithms.

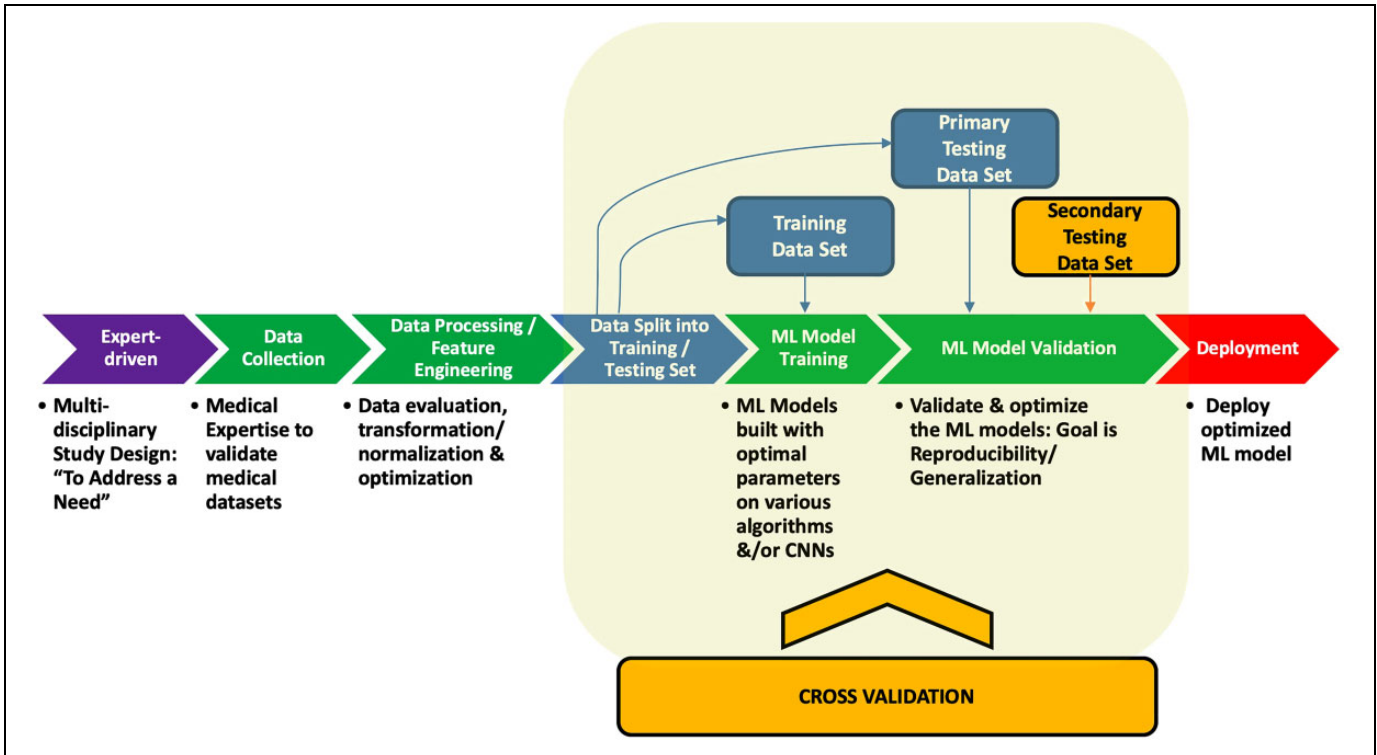
#### 4) Are the optimized ML models applicable and generalizable when applied to a novel data set?

Selecting the appropriate algorithm is essential in finding the most suitable model for a given task. Hence, to enhance the algorithm’s predictive capability (most importantly its ability to generalize), an optimal study design along with an iterative validation process is required.

### Supervised Machine Learning Study Design and Validation

After data are collected, cleaned, and preprocessed, and the correct ML approach has been chosen, the next step is model building and validation studies which ultimately yields the deployed model (Figure 4). The supervised ML model building phase usually includes splitting the data into an initial training and testing set that allows training of the model followed by testing for its initial validation phase. To minimize overfitting of the models, certain model adjustments and incorporating cross-validation (CV) processes allows the empirical build of a large number of models whose performances can be subsequently assessed with the goal of finding the most generalizable model. It is well known that assessment based on the initial validation test set does not always yield a generalizable model as we have shown in our recent studies.<sup>20</sup> Hence, it is essential to include secondary and sometimes tertiary external test sets (previously unseen by the model) to assess its true generalizability.

In brief, the model is initially trained and preliminarily validated on its train-test split data set. An example of this is where an “80-20 train-test” split initially trains the model using 80% of the data, followed by the remaining 20% which is used for testing and its initial validation (Figure 4). However, this approach alone for building a single ML model is prone to overfitting. Hence, to minimize the overfitting phenomena as noted earlier, good practice demands that one build large number of models with variable parameters through one or more CV platforms. Some of the most commonly employed CV



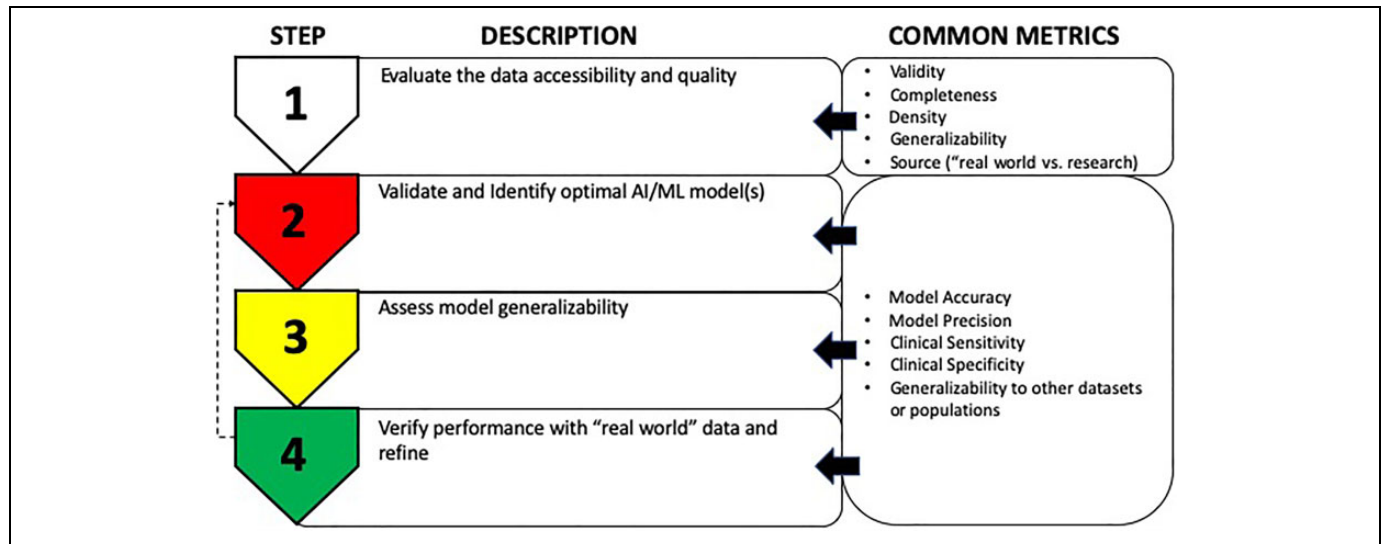
**Figure 4.** Supervised (labeled) machine learning model study design overview. Steps for the deployment of a supervised machine learning model. From left to right, the figure shows the initial team of multidisciplinary experts defining a study design to address a need. Data are then collected, processed, trained tested, validated, and ultimately deployed.

studies include the “*k*-fold” CV, “leave-one-out” CV, and bootstrapping. In *k*-fold CV, the train test data set is split *k* times. For instance, if the *k* is 10, the data are split into 10 train-test splits to assure proper sampling of the training set and more importantly the testing data. This approach assures a better sampling for the test sets and minimizes selection bias which ideally leads to a more generalizable model as we have demonstrated in a recent study.<sup>11</sup> Leave-one-out CV is a similar concept but is the extreme version of the *k*-fold approach in which *k* will equal to *n* (eg, total number patients being studied). Instead of train-test splitting them into *k*-folds of 5 or 10, as the name implies, *k* will equal the total number of individual data entries *n* (eg, *n* of 100 could be the individual sets of data in 100 patients). In each train-test run, one (eg, one patient data set) is left out for testing phase which leads to a complete sampling of the data set, but this approach may not always enhance the model dramatically and can also be computationally very demanding. In contrast to *k*-fold and leave-one-out CV, in bootstrapping one creates a new data set with the same size as the original data set by randomly pulling samples from the original data set. As is evident, this method may result in duplicate data being used in the new bootstrapped data set. This method is commonly employed in certain ensemble tree algorithms such as RF in which a random subset of the bootstrapped data set is used in creating the decision trees. A bootstrapped data set that uses the aggregated data to make a decision is called “Bagging” which stands for bootstrapping aggregation.

In this approach, a proportion of data within this randomly selected bootstrapped data set is not present (Out of Bag) which can be subsequently used to test the trained model within the bootstrapped data set to assess the accuracy of the model. The use of such CV approaches within the algorithm’s building phase can ultimately help in finding suitable models that can then be secondarily tested on a separate data set to assess their true generalizability potential.

## Summary

Artificial intelligence and ML have the potential to transform health care in the coming years. To ensure that pathologists and laboratorians are equipped to play important roles in the multidisciplinary teams, we have provided definitions, descriptions, and an outline of 4 of the essential steps for developing AI/ML applications. The need for high-quality data (Figure 5) illustrates the role of pathologists and laboratorians in appropriately curating, interpreting, and providing results for AI/ML applications. We encourage a balanced approach utilizing clinical trial data, when available, combined with real-world data to optimize AI/ML training. The approach and technique chosen should be tailored to the data available and the problem to be solved. Since many AI/ML techniques are available and not all are the same, pathologists and laboratorians must be sufficiently familiar and literate with these options so that they can communicate effectively and make meaningful contributions



**Figure 5.** Stepwise considerations for development and validation of the machine learning (ML) model. The figure describes a very general stepwise approach for development and validation of an ML model. Common metrics used in each step are shown on the right. Step 1 involves assessing the quality and accessibility of the data, followed by step 2 that requires method validation to identify optimal ML model(s). Once optimal ML models have been identified, step 3 involves determining their ability to work with other data sets to assess generalizability. Finally, step 4 involves evaluating the data in more "real-world" conditions to further assess performance and generalizability along with further refinement (go back to step 2) to improve the performance and desirable outcomes.

within the AI development team. Determining the overall generalizability of AI/ML models for real-world populations is critical to most successful development and implementation strategies. Researchers in this area are encouraged to be aware of their data limitations and develop cross-disciplinary literacy in AI/ML methods to effectively harness their optimal implementation plan, thus maximizing its impact.

### Acknowledgments

Special thanks to all our machine learning collaborators who keep us energized in this exciting new arena.


### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### ORCID iD

Elham Vali Betts  <https://orcid.org/0000-0002-0773-5136>

### References

- EMC Digital Universe. *IDC Vertical Industry Brief: The digital universe driving data growth in health care; challenges and opportunities for it 2014*. Vertical Industry Brief. <https://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pdf>. Accessed December 29, 2015.
- Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev*. 1959;3:210-229.
- Koza JR, Bennett FH III, Andre D, Keane MA. Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In: Gero JS, Fay S, eds. *Artificial Intelligence in Design '96*. Berlin, Germany: Springer; 1996:151-170.
- The Center for Devices and Radiological Health (CDRH). *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)—Discussion Paper and Request for Feedback*. In: Food and Drug Administration (FDA). 2019.
- Becich MJ. Information management: moving from test results to clinical information. *Clin Leadersh Manag Rev*. 2000;14:296-300.
- Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med*. 2018;15:e1002689-e1002689.
- Christenson RH, Jacobs E, Uettwiller-Geiger D, et al. Comparison of 13 commercially available cardiac troponin assays in a multicenter north American study. *J Appl Lab Med: An AACC Publication*. 2017;1:544-561.
- Lee CH, Yoon HJ. Medical big data: promise and challenges. *Kidney Res Clin Pract*. 2017;36:3-11.
- Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and impact of real-world clinical data for the practicing clinician. *Adv Ther*. 2018;35:1763-1774.
- Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*. 1-59593-383-2. Pittsburgh, PA; 2006:161-168.

11. Tran NK, Sen S, Palmieri TL, et al. Artificial intelligence and machine learning for predicting acute kidney injury in severely burned patients: a proof of concept. *Burns*. 2019;pii: S0305-4179(18)31129-X.
12. Buehler L, Rashidi HH. *Bioinformatics Basics, Application in Biological Science and Medicine*. 2nd ed. CRC Press, Taylor and Francis Group; 2005.
13. Chapman P, Clinton J, Kerber R, et al. *CRISP-DM 1.0 Step-by-Step Data Mining Guide*. CRISP-DM; 2000. <https://www.the-modeling-agency.com/crisp-dm.pdf>. Accessed February 2019.
14. Shearer C. The CRISP-DM model: the new blueprint for data mining. *J Data Warehouse*. 2000;5:13-22.
15. Aloise D, Deshpande A, Hansen P, Popat P. NP-hardness of Euclidean sum-of-squares clustering". *Machine Learning*. 2009; 75:245-248.
16. Cordeiro A, Boris MR. Minkowski metric, feature weighting and anomalous cluster initialisation in k-means clustering. *Pattern Recognition*. 2012;45:1061-1075.
17. Celebi ME, Kingravi HA, Vela PA. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Syst Appl*. 2013;40:200-210.
18. Greenmeier L. *20 Years After Deep Blue: How AI has Advanced Since Conquering Chess*. Scientific American; 2017. <https://www.scientificamerican.com/article/20-years-after-deep-blue-how-ai-has-advanced-since-conquering-chess>. Accessed February 2019.
19. Murphy M. *Google's AI Just Cracked the Game that Supposedly No Computer Could Beat*. Quartz; 2016. <https://qz.com/603313/google-ai-just-cracked-the-game-that-supposedly-no-computer-could-beat>. Accessed February 2019.
20. Jones AD, Graff JP, Darrow M, et al. Impact of pre-analytic variables on deep learning accuracy in histopathology. *Histopathology*. 2019;75:39-53.
21. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. *Neural Computation*. 1992;4:1-58.
22. James GM. Variance and bias for general loss functions. *Mach Learn*. 2003;51:115-135.
23. Valentini G, Dietterich TG. Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. *J Mach Learn Res*. 2004;5:725-775.
24. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Salmon Tower, NY: Springer-Verlag; 2013.
25. Wahab L, Jiang H. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. *PLoS One*. 2019;14:e0214966.
26. Seal HL. Studies in the history of probability and statistics. XV The historical development of the Gauss linear model. *Biometrika*. 1967;54:1-24.
27. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. *Biometrika*. 1967; 54:167-179.
28. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med*. 1997;16:965-980.
29. Palei SK, Kumar Das S. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: an approach. 2009;47:88-96.
30. Biondo S, Ramos E, Deiros M, et al. Prognostic factors for mortality in left colonic peritonitis: a new scoring system. *J Am Coll Surg*. 2000;191:635-642.
31. Ranganathan P, Aggarwal R, Pramesh C. Common pitfalls in statistical analysis: odds versus risk. *Perspect Clin Res*. 2015;6: 222-224.
32. Ranganathan P, Pramesh C, Aggarwal R. Common pitfalls in statistical analysis: logistic regression. *Perspect Clin Res*. 2017; 8:148-151.
33. George JH, Langley P. Estimating continuous distributions in Bayesian classifiers. *Paper presented at: Eleventh Conference on Uncertainty in Artificial Intelligence*; August 18–20, 1995; Montréal, Qué, Canada.
34. Hand DJ, Yu K. Idiot's Bayes: Not So Stupid After All? *Int Stat Rev/Revue Internationale de Statistique*. 2001;69:385-398.
35. Rish I. An empirical study of the naive Bayes classifier. Paper presented at: *IJCAI Workshop on Empirical Methods in AI*; August 4–10, 2001; New York, NY.
36. Hyafil L, Rivest RL. Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*. 1976;5:15-17.
37. Quinlan JR. Induction of decision trees. *Machine Learning*. 1986; 1:81-106.
38. Papagelis A, Kalles D. Breeding decision trees using evolutionary techniques. *Proceedings of the Eighteenth International Conference on Machine Learning*; June 28–July 1, 2001; San Francisco, CA, USA.
39. Mehta D, Raghavan V. Decision tree approximations of Boolean functions. *Theor Comput Sci*. 2002;270:609-623.
40. Friedman JH, Meulman JJ. Multiple additive regression trees with application in epidemiology. *Stat Med*. 2003;22:1365-1381.
41. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *J Anim Ecol*. 2008;77:802-813.
42. Coomans D, Massart DL. Alternative k-nearest neighbour rules in supervised pattern recognition: part 1. k-Nearest neighbour classification by using alternative voting rules. *Analytica Chimica Acta*. 1982;136:15-27.
43. Altman NS. An introduction to kernel and nearest-neighbor non-parametric regression. *The American Statistician*. 1992;46: 175-185.
44. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. *When Is "Nearest Neighbor" Meaningful?* Berlin, Heidelberg, Germany: Springer-Verlag; 1999.
45. Nigsch F, Bender A, van Buuren B, Tissen J, Nigsch E, Mitchell JB. Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. *J Chem Inf Model*. 2006;46:2412-2422.
46. Peter H, Park BU, Samworth RJ. Choice of neighbor order in nearest-neighbor classification. *Ann Stat*. 2008;36:2135-2152.
47. Hearst MA. Support vector machines. *IEEE Intell Syst*. 1998;13: 18-28.
48. Cho G, Yim J, Choi Y, Ko J, Lee SH. Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investig*. 2019;16:262-269.
49. Andy L, Wiener M. Classification and regression by randomForest. *R News*. 2002;2: 18-22.



50. Tin Kam H. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell.* 1998;20: 832-844.
51. Breiman L. Random forests. *Mach Learn.* 2001;45:5-32.
52. Shi T, Seligson D, Belldegrin AS, Palotie A, Horvath S. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Mod Pathol.* 2005;18: 547-557.
53. Lin Y, Jeon Y. Random forests and adaptive nearest neighbors. *J Am Stat Assoc.* 2006;101:578-590.
54. Prinzie A, Van den Poel D. Random Forests for multiclass classification: random multinomial logit. *Expert Syst Appl.* 2008;34: 1721-1732.
55. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med.* 2018;24:1559-1567.