Seminars in Diagnostic Pathology 40 (2023) 71–87

Contents lists available at ScienceDirect

Seminars in Diagnostic Pathology

journal homepage: www.elsevier.com/locate/semdp

Review article

Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts

Samer Albahra [a,c,*], Tom Gorbett [a,c], Scott Robertson [a,c], Giana D'Aleo [a,c], Sushasree Vasudevan Suseel Kumar [a,c], Samuel Ockunzzi [a,c], Daniel Lallo [a,c], Bo Hu [b,c], Hooman H. Rashidi [a,c,*]

[a] Pathology and Laboratory Medicine Institute (PLMI), Cleveland Clinic, Cleveland, OH, United States
[b] Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, United States
[c] PLMI's Center for Artificial Intelligence & Data Science, Cleveland Clinic, Cleveland, OH, United States

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning
Pathology
Laboratory medicine
Supervised
Learning
Predictive modeling

ABSTRACT

Machine learning (ML) is becoming an integral aspect of several domains in medicine. Yet, most pathologists and laboratory professionals remain unfamiliar with such tools and are unprepared for their inevitable integration. To bridge this knowledge gap, we present an overview of key elements within this emerging data science discipline. First, we will cover general, well-established concepts within ML, such as data type concepts, data preprocessing methods, and ML study design. We will describe common supervised and unsupervised learning algorithms and their associated common machine learning terms (provided within a comprehensive glossary of terms that are discussed within this review). Overall, this review will offer a broad overview of the key concepts and algorithms in machine learning, with a focus on pathology and laboratory medicine. The objective is to provide an updated useful reference for those new to this field or those who require a refresher.

Introduction

The practice of medicine generates large volumes of complex data requiring the need for advanced analytical tools. More and more, these tools are becoming centered about machine learning. Machine learning is within the umbrella of artificial intelligence (AI) and incorporates numerous key elements from both statistics and computer science disciplines. The goal within most machine learning tasks is to find patterns within the data that can then generate models capable of making predictions on new unforeseen data.

Pathology and laboratory medicine play central roles in many medical decisions whose complex and ever-growing data are increasingly in need of machine learning integration. Additionally, the role of the pathologists and laboratory professionals as gatekeepers of such data further supports the need for them to become increasingly familiar with the capabilities and known limitations within this space. As professionals who are well versed in best practices of test development, it is logical for us to also help guide the development of ML tools since they share many similarities with laboratory-developed tests (LDTs). Even though there are key similarities between LDTs and ML testing platforms, there are also unique differences within each that further add to the challenges that we face as we deploy such tools. One major difference revolves around the changing data that is inevitable within the ML studies with time (especially for the tabular data tasks). In essence, this is referring to the concept of data drift which is definitional for most tabular data ML studies but not a major point of discussion for most LDTs. Regardless of the differences and challenges, the need for standardization of key aspects of these advanced analytics platforms (i.e., study design and machine learning operations) should remain a high priority that will not only ensure their reproducibility but also highlight their true value within the medical arena. The absence of such integral processes and standards can potentially compromise our patient care efforts as we embrace these tools in our settings. Additionally, by carefully monitoring the ML models after deployment, these measures will ultimately ensure the quality of the data and the model's predictive outcomes. A general understanding of these essential domains is critical

* Corresponding authors at: 9500 Euclid Ave, Cleveland, OH 44195.
E-mail addresses: albahrs@ccf.org (S. Albahra), rashidh@ccf.org (H.H. Rashidi).

https://doi.org/10.1053/j.semdp.2023.02.002

Available online 16 February 2023

for their integration and best use. Therefore, bridging the knowledge gap in this space for all involved will not only be instrumental for their enhanced success rates but more importantly will serve as a critical step for continuously improving the quality of these studies and their associated implemented ML models.

To that end, the main purpose of this article is to build on the foundation of prior excellent AI/ML review articles and to provide the readers with updated information on specific aspects of the machine learning arena. Specifically, the article will focus on data types, data preprocessing tasks, and how each of these feeds into the various machine learning algorithms employed along with their common ML-associated terminologies (provided within a relatively comprehensive glossary of terms; Table 1).

Data types

Everything in machine learning starts with the data and the various data types drive the respective machine-learning approaches that are employed within this process. The main four data types involved include image, text, audio, and tabular/numerical data.

1a. Image data: This data type consists of visual information, which is typically organized as arrays of pixel values. An example of this in pathology could be microscopic images (from a whole slide image: WSI) from various tissue samples (e.g., colon cancer versus normal colon). These could then be used to train certain ML models that can ultimately help differentiate various images (e.g., to distinguish cancer versus normal tissue).

For these to become ingestible within machine learning libraries, the image data will typically need to be organized into multidimensional arrays or tensors, each consisting of various image characteristics (e.g., height and width of the image, along with a variety of other intrinsic features within the image). There could also be metadata attached to these files which could represent a variety of other associated specific image features (e.g., video frame number, specific location data, etc.).

1b. Text data: These consists of individual words or sentences that can be employed within the various natural language processing (NLP) platforms. NLP ultimately enables computers to read, understand, and generate human language. After the tragedies of World War II, globalization became a priority, and NLP was born in the hopes of automating translation from one spoken language to another[2]. Like any other non-numerical data type, text data will also need to be transformed into a numerical representation for them to become machine compatible. Examples within diagnostic surgical pathology could include NLP integration for transcription or search tasks. With studies dating back to the 1950s and 60s, NLP has always contained its own set of unique challenges[3]. Classic challenges within NLP include the need for addressing ambiguities (such as homophones), syntax and semantics (i.e., context) issues, dealing with anaphora, and pragmatics awareness to list a few. Homophones, for instance, are difficult for machines to understand, especially if we consider each word without context[4]. If I am looking at a sentence that contains the word "bark" without any knowledge of the discussion that is happening around the said word, I wouldn't know if we were talking about dogs or trees, or misspellings of 18th century sailing vessels[5]. Another important challenge in NLP is understanding context[6]. Whether we realize it or not, our minds have developed a deep knowledge of how we use language in numerous scenarios and how these culminate across different spoken languages[7]. Machines do not have this advantage, so they have to take into account corpora[8]. These refer to a large collection of text data that can be used to train and evaluate the NLP algorithms which will be discussed in more detail in the pre-processing section. Many complex subtleties in language are hard to translate into code, which makes context crucial and further complicates the use of NLP models. However, newer NLP-based platforms and approaches (such as OpenAI's ChatGPT) are starting to address many of these key issues and scratching the surface of Artificial General Intelligence (see Table 1 for more details).

1c. Audio data: This form of data is comprised of sound waves. The classic example in medicine involves heart sounds, which the ML model can be trained on to discover certain underlying medical conditions. This type of data is not routinely used within pathology and laboratory medicine since most information within our field involves image, text, and tabular data formats.

1d. Tabular/Numerical data: Numerical data consist of measurements and counts for various clinical, laboratory, or historical values and are the most abundant data present within the healthcare space. The correct representation of numerical data will ensure that the findings are systematic and reproducible. These are typically categorized as qualitative or categorical (i.e., nominal, and ordinal data types) versus quantitative values (interval or ratio data types).

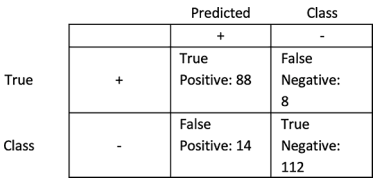
Qualitative data (categorical data) answers the questions revolving around "What/Which category?" (e.g., abnormality levels of a test result: normal versus high versus low and cancer versus normal tissue, to name a few). As noted, these can also be further grouped into nominal and ordinal data types[9,10]. Nominal data does not follow any order between the categories (e.g., eye color category of brown, blue, or green) while ordinal data can be ranked or ordered (e.g., high-grade, intermediate-grade, and low-grade tumors). Within these categorical/qualitative data types, if the number of options is confined to just two values (e.g., cancer versus normal), then these are better known as binary categories.

In contrast to qualitative data, quantitative data answers the question "How much/many?" (e.g., number of pregnancies, patient hemoglobin values, BMI, etc.). These can be further classified into either discrete classes or continuous numerical data or as interval data (no zero point present as in BMI) versus ratio (has a true zero point as in the number of pregnancies) data types[9,10]. In summary, continuous data represents any value within a certain range and can include decimals (e.g. body temp 98.1F) as opposed to discrete data that can only take on a limited set of values (e.g. the number of medications a patient is taking is 3)[11].

Many of these data subtypes are more commonly seen within laboratory medicine and clinical disciplines rather than surgical pathology since these values from a patient's electronic medical record (EMR) are typically extracted in tabular formats.

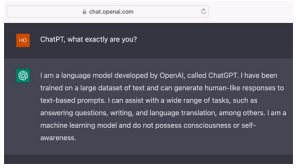
Regardless of which data is present, many tabular data files contain a combination of the aforementioned data subtypes which are used to train their respective machine learning algorithms for various tasks (e.g., prediction of cancer or sepsis)[12]. These values are typically presented within organized rows and columns within the given table. The values within each row generally represent an individual (i.e., patient instance) and each column represents the various variables (i.e., features/independent variables or target/dependent variable) for a given patient. The features (independent variables) within the tabular dataset are ultimately mapped (through some function) to the target of interest (dependent variables/target or outcome) to acquire the ML model that can be used for making future predictions on unforeseen new data. For a supervised ML model these can also be represented as follows: $Y = f(X) + e$ where Y represents the target of interest, X represents the features, f represents the function (the acquired mathematical relationship between the X and Y) and e representing the irreducible error since no ML model is ever perfect. However, before proceeding to the machine learning algorithms and training steps to map the features to the target, the data of interest will likely require various cleaning and standardization tasks which are collectively known as data pre-processing.

Table 1
Glossary of terms.

Term	Definition	Example from medicine
Accuracy	A statistical metric used to evaluate the performance of an ML model. It represents the number of correct predictions (true positives and true negative cases) over all predictions (true positives TP, true negatives TN, false positives FP and false negative FN cases). Accuracy = True predictions / All predictions = (TP+TN) / (TP+TN+FP+FN) Note: Sometimes the terms "Accuracy" and "Precision" are confused. Accuracy shows how close a measurement is to the true value while Precision signifies how close measurements of the same item are to each other (see the term "precision" below for more details).	In the following example, this cancer prediction model has resulted in 88 True positive cases and 112 True Negative cases. These are positive cases in real life that the model is predicting correctly as positive and negative. Accuracy = (112 + 88) / (112 + 88 + 8 + 14) = 90% 
Activation Functions	A mathematical function used in an artificial neural network that defines the output state of a node based on the input(s)	In a medical imaging application, a sigmoid activation function could be used to classify whether a tumor is benign or malignant based on its image characteristics
Algorithm	A method (i.e., set of instructions) that a computer follows to accomplish a task. Some common algorithms in supervised machine learning include Logistic regression, neural networks, K-NN, random forest, support vector machine, and gradient boosting machine.	A logistic regression algorithm that has been trained to distinguish sepsis from no sepsis cases.
Anomaly Detection	AKA outlier detection, is a method that can identify data points that are unusual or unexpected (i.e., deviate significantly from the data majority).	May be used to identify outlier vital sign measurements compared to normal reference ranges (e.g., very high heart rates or temp).
Artificial General Intelligence (AGI)	AI that is on par with humans and capable of general intelligence (i.e., can think like humans). It is mainly in the research realm and considered to be in the "strong AI" domain (like Artificial Super Intelligence or ASI which theoretically surpasses humans and currently does not exist)	Although ChatGPT is technically an Artificial Narrow Intelligence, it is starting to scratch the surface of AGI
Artificial Intelligence	Paraphrasing Arthur Samuel and others, it is the capability of machines to imitate intelligent human behavior. This umbrella term contains other specific disciplines such as Artificial Narrow Intelligence which includes machine learning (ML) and Artificial General Intelligence	ML is an example of Artificial Narrow Intelligence (generally placed in the Weak AI category) while Artificial General Intelligence (AGI) is a Strong AI category
Artificial Narrow Intelligence (ANI)	These are collectively considered to be in the "weak AI" domain and include our traditional ML methods along with the more advanced ones such as deep learning. These are designed to perform specific tasks (generally better than humans) and automate many processes. However, they are not capable of general intelligence and can only perform their specific designed task	Examples include many of the ML-based models such as Siri, Alexa, Image classification models (that distinguish cancer from non-cancerous tissue) and various tabular data classifier models (e.g., random forest sepsis predictor), etc.
Automated Machine Learning (Auto-ML)	A process that can automatically select and optimize the best ML model for a given data with no or little human intervention.	MILO Auto-ML is used to automatically find the best performing model for a sepsis study (i.e., best model that can distinguish sepsis from non-septic patients)
Batch Size	In deep learning, these are the batched number of samples used within the dataset in one iteration of training. This specified hyperparameter allows users to control the number of training examples that are processed at once by the model (notably higher batch sizes will inevitably require more memory usage, therefore the optimal batch size is memory dependent)	When using TensorFlow to build a deep learning model to classify normal from abnormal histology images (e.g., colon cancer versus normal colon histology), we can set batch size to 64 which signifies using 64 images at a time to train the model
Bagging (Bootstrap Aggregating)	Through this approach we create multiple versions of the same model (each trained on a different subsample of the original dataset) who's averaged predictions yield the model's final prediction. This method is commonly combined with other ensemble ML algorithms (e.g., random forest) to improve their performance.	A random forest algorithm that employs bagging is used to find an optimized ML model that distinguishes caved from non-covid cases.
Bayes Theorem	The formula that describes the relationship between the conditional probabilities of events, and provides a method for calculating the posterior probability of an event given certain evidence	The probability (P) of pneumonia (A) given the presence of "fever" (B) $P(A\|B) = P(B\|A) * P(A) / P(B)$
Bias-Variance Tradeoff	The tradeoff between the error introduced by the bias of a model and the error introduced by the variance in the model's predictions. A model with low bias and high variance may overfit the data and perform poorly on new data, while a model with high bias and low variance may underfit the data and perform poorly. Finding the right balance between these two is critical for optimizing the model's performance.	A covid prediction ML model was deployed which showed a great performance on the initial training and validation data, but this ML model performed poorly on the follow up external data. The investigators concluded that this was due to overfitting which explained the model's lack of generalizability.
Bit	Represents the two possible states (either 0 or 1) of a digital signal	In medical digital imaging each pixel may be represented by 8 or more bits (i.e., bytes)
Boosting	An ML technique that is typically used in some ensemble methods (e.g., Gradient Boosting Machine; GBM) to combine the performance of multiple weak models into a single strong model.	A GBM algorithm is used to make a tuberculosis prediction model by combining multiple weak learners (i.e., trees) to create a more accurate ensemble classifier
Bootstrapping	This is a statistical method that makes estimates of the distribution of a sample statistic by resampling the data with	Bootstrapping was used as part of the bagging method in the random forest model to make a sepsis versus no sepsis ML model

(*continued on next page*)

Table 1 (*continued*)

Term	Definition	Example from medicine
	replacement. These samples are drawn with replacement from the original dataset to form a new set of samples (e.g., used to estimate the population's distribution)	
Brier Score	Measures the accuracy of probabilistic predictions. It ranges from 0 to 1, with 0 indicating perfect accuracy and 1 indicating total inaccuracy. It is calculated as the mean square error between predicated probabilities and the actual outcome.	The Brier score is evaluated when comparing similar COVID prediction models to find the model with the most reliable probability score (lowest Brier score).
ChatGPT	This pre-trained language model is a type of AI called "Generative Pre-trained Transformer" (GPT). GPT models are trained on large text data and capable of generating coherent, contextually appropriate responses to text-based prompts. Specifically, it is a Transformer-based neural network model which uses an attention mechanism to weigh the importance of different parts of the input when generating the output. This attention mechanism allows the model to focus on specific parts of the input when generating text, which helps to create more coherent and contextually appropriate output. Almost like a real person. Although this is still technically considered a "narrow" AI (ANI), it is starting to scratch the surface of AGI (Artificial General Intelligence).	Here's an example of ChatGPT in action: 
Classification	The process of assigning or categorizing a given set of data into specific categories or classes	A classification algorithm was used to identify tumor types based on their characteristics on a medical image
Clustering	The process of grouping data points into "clusters" based on their shared similarities (i.e., the similarity between the data in a cluster is greater than its similarities to other clusters). This is typically an unsupervised method within machine learning realm	k-means method (a popular clustering algorithm) was used in a gene expression study, which identified different subtypes of a disease
Convolutional neural network	A type of artificial neural network that is particularly effective at analyzing image data	In pathology, convolutional neural networks can be used to analyze medical images for tasks such as tumor detection (cancer vs normal), object detection (identifying neutrophils and lymphocytes within the same image) or segmentation.
Confidence interval (CI)	A range of values that is likely to contain the true value of a population parameter with a certain level of confidence. In short, A range of values that is likely to contain the true value of a parameter	Two acute kidney injury prediction models trained and validated on two different sized datasets showed similar accuracies (e.g., 91%) but different 95% confidence intervals (Model A had a 95% CI of 89%−92% while Model B's 95% CI was 82%−94%). Hence, the narrower 95% CI (89%−92%) is the more reliable model (model A) and most likely included the larger dataset size as well.
Confusion matrix	A table that is often used to describe the performance of a classification model, comparing the model's predictions to the actual targets. These can be binary or multiclass. The simplest confusion matrix is based on a binary classifier and is a 2 × 2 table (to display the true positive, true negative, false positive and false negative results) while those for multi classification tasks have more elaborate tables to display these results	This binary confusion matrix was used to evaluate the accuracy, sensitivity, specificity, and many other performance measures of this sepsis prediction model (yielding the following results) Predicted Class True Class: + / True Positive: 88; False Negative: 8 True Class: - / False Positive: 14; True Negative: 112
Correlation	A measure of the strength and direction of the relationship between two variables	For example, the hemoglobin and hematocrit values in a dataset may be 98% correlated while hemoglobin and the temperature values in the dataset may be 51% correlated. This displays how hemoglobin and hematocrit are closer related than hemoglobin and temperature are (at least within this hypothetical dataset)
Cropping	The process of removing unwanted parts of an image or data set to focus on a specific region of interest	In medical imaging, cropping can be used to remove non-essential parts of an image, such as the background, to improve the accuracy of image analysis
Cross Validation (CV)	A method that uses different data subsets to evaluate the performance of the ML model (this can help minimize overfitted models)	The classic example is k-fold cv in which the data is divided into k subsets (e.g., 5) which means that there will be 5 variations of the training and testing data as the data is split into these groups (hence, in this case tested on 5 different test subsets).
Decision tree	These ML algorithms use a tree-like model to make predictions based on the characteristics of the data and can be used for both classification and regression tasks.	Simple decision trees are not commonly employed for our ML studies. However, the ensemble tree methods (e.g., random forest and Gradient boosting machine) are found to be very capable in providing high performing ML models (e.g., sepsis versus no sepsis prediction)
Dependent variable	These are also referred to as the "target" (sometimes also called outcomes column). In a supervised method ($Y = f(X) + e$), these represent the Y in which the features (independent variables X) are mapped to them through a function f. In summary, this is the variable that is predicted by a machine learning model based on the input data.	A simple binary dependent variable (i.e., target) could be the presence or absence of cancer. For example, the ML model ultimately will be able to make a distinction between cancer cases and non-cancer cases based on histologic features that are present within the images.
Ensemble method	A machine learning technique that combines the predictions of multiple models to produce a more accurate or stable prediction	Random forest and Gradient boosting machine are some classic ensemble tree methods capable in providing high performing ML models (e.g., sepsis versus no sepsis prediction)
Epoch	This refers to a single pass through the entire training data. In other words, a single iteration in a machine learning model's training process, during which the model is presented with all the training data. This is typically one of several hyperparameters that can be set within a neural network model.	It is important to understand the relationship between epoch, batch size and iteration parameters in a ML model training process. Number of iterations = (total number of epochs * (total number of examples / batch size)) For example, if we had 4000 samples in our training step, then these 4000 samples can be divided into batches of 500, which means that 1 epoch will take 8 iterations to complete

(*continued on next page*)

Table 1 (*continued*)

Term	Definition	Example from medicine
F1	In classification models, the F1 score represents the harmonic mean of precision (i.e., positive predictive value) and recall (i.e., sensitivity) of the model and defined as: F1 = 2 * (precision * recall) / (precision + recall)	In real-world breast cancer diagnosis scenario, a balance between precision (positive predictive value) and recall (sensitivity) of the ML model need to be achieved. These two measures typically move in opposite directions (i.e., when sensitivity goes up precision usually goes down and vice versa). This allows us to find the balance between our false positive and false negative cases since false negatives can sometimes severely impact patient's life (e.g., in a cancer prediction ML model), while a false positive can lead to unnecessary tests and/or treatments.
False Negative	In a classification model, these represent the positive cases in real life that the model incorrectly predicts as negative	In the following example, this cancer prediction model has resulted in 8 false negative cases. These are positive cases in real life that the model is predicting incorrectly as negative cases. Predicted Class + | - True + | True Positive: 88 | False Negative: 8 Class - | False Positive: 14 | True Negative: 112
False Positive	In a classification model, these represent the negative cases in real life that the model incorrectly predicts as positive. False positive rate (FPR) and specificity are inversely proportional. Hence, in an ROC curve, the X axis of the curve represents the FPR as 1-specificity. (See ROC for more details)	In the following example, this cancer prediction model has resulted in 14 false positive cases. These are negative cases in real life that the model is predicting incorrectly as positive cases. Predicted Class + | - True + | True Positive: 88 | False Negative: 8 Class - | False Positive: 14 | True Negative: 112
Features	These are also referred to as the "independent variables" (sometimes also called predictors or attributes). In a supervised method ($Y = f(X) + e$), these represent the X in the formula above that map to the target Y through a function f.	For example, in our ML model, the following "features" (heart rate, temperature, hemoglobin and ECG findings) are mapped to the heart disease presence or absence target column. Therefore, values for the above features from future patients can be inserted into the ML model to tell us if they are positive or negative for heart disease.
GAN (Generative Adversarial Network)	A type of machine learning algorithm that involves two competing neural networks (hence the term "Adversarial"), known as the generator and the discriminator. The generator attempts to produce data that is like a training dataset, while the discriminator tries to distinguish real data from the generated data.	A GAN could be used in medical imaging to generate synthetic CT scans or MRI images that can be used for training other machine learning algorithms. It can also be used to build synthetic tabular/numerical data to create synthetic heart disease data from the real data counterpart
Generalizability	The ability of a machine learning model to make accurate predictions not only on its initial training/validation test but also on new, unseen data (secondary or tertiary datasets).	Our prostate cancer model was generalizable since it had very similar performance measures (sensitivity, specificity, accuracy, F1, etc.) on both the initial validation test and the follow up secondary and tertiary test sets. In contrast, our colon cancer model was shown to be overfitted, which meant that it performed well on the initial validation test but poorly on the follow up secondary and tertiary test sets (i.e., not generalizable).
Gradient Descent	This is an optimization method used to minimize a function (typically the cost function of a machine learning model). It is an iterative process that starts with an initial guess for the parameters of the model and updates the parameters in the direction of the steepest descent of the cost function with respect to the parameters, until it reaches a local or global minimum.	Gradient Descent is often used in deep learning such as in building an ML model that can differentiate prostate cancer cases from their non-cancerous counterparts. It's also worth mentioning that deep learning models have multiple hyperparameters, and the optimization can be computationally expensive, which is why variations of the gradient descent approach may be employed (e.g., Adam) to optimize these.
Hyperparameters	These model parameters are set before training a machine learning model, as opposed to parameters that are learned during training. They help control the learning process and the capacity of the model to learn. "Hyper" in hyperparameter refers to the fact that they are "higher-level" (not learned from the data during training) as opposed to the regular parameters of the model which are learned from the data during training.	The following are some common examples of hyperparameters in various ML algorithms: -Number of hidden layers and number of neurons in each layer in a neural network -Number of trees and the maximum depth of each tree in a random forest -Regularization strength: The degree to which the model is regularized to prevent overfitting. -Number of neighbors measured in K-NN
Imputation	The process of replacing missing values in a dataset with some value based on inference	Examples include: Mean imputation: replace the missing values with the mean of the values Regression imputation: uses a predictive model, such as linear regression, to predict the missing values based on the observed values of other variables Multiple Imputation, etc.
Independent variable	These are also referred to as the "features" (sometimes also called predictors or attributes). In a supervised method ($Y = f(X) + e$), these represent the X in the formula above that map to the target Y through a function f.	For example, in our ML model, the following "independent variables/features" (heart rate, temperature, hemoglobin and ECG findings) are mapped to the heart disease presence or absence target column. Therefore, values for the above features from future patients can be inserted into the ML model to tell us if they are positive or negative for heart disease.
Irreducible error	In a supervised method ($Y = f(X) + e$), X represents the features that are mapped to the target Y through a function f. "e" represents the Irreducible error. This is the error that cannot be reduced or eliminated by any machine learning algorithm, due to the inherent noise or variability in the data	In medicine, irreducible error could arise from factors such as measurement error or inter-observer variability

(*continued on next page*)

Table 1 (*continued*)

Term	Definition	Example from medicine
Iterations	Number of times a machine learning algorithm is applied to the data during training. During each iteration, the model processes a batch of training examples, which are a subset of the entire training dataset Number of iterations = (total number of epochs * (total number of examples / batch size))	For example, if the training dataset contains 1000 examples, and the batch size is set to 100, then each iteration would process 100 examples, and the training process would require 10 iterations to process the entire training dataset.
k-fold	A method of cross-validation in which the data is divided into k equal-sized subsets, and the algorithm is trained and evaluated k times, each time using a different subset as the test set	For example, a k-fold of 10 means that an 80%−20% train-test split data will have 10 versions of this train and test sets (i.e., different 20% test sets that will better represent the model's true performance on the entire data evaluated). This helps minimize overfitting (especially true on big data studies).
K-Means	A type of clustering algorithm (unsupervised ML approach) that groups data points into k clusters based on their similarities	In medicine, K-Means can be used to identify subgroups of patients with similar symptoms or treatment responses
K-NN	A type of non-parametric machine learning algorithm (for data classification and regression tasks) that uses the k number of training examples that are closest to a new input, and then uses the most common class among these k nearest neighbors as the prediction for the new input	K-NN model was used to build an acute kidney injury prediction model within the burn's population.
Kernel Function	a function used to transform the data into a higher dimensional space to become easier to separate and to find a decision boundary	A kernel function is commonly used in support vector machine (SVM) and typical kernels include linear, RBF, sigmoid, and polynomial
Keyword Extraction	A natural language processing (NLP) technique for identifying the most important words or phrases in a document	Using keyword extraction to identify the most important symptoms in a patient's medical records.
Leave-one-out	A method of cross-validation where a model is trained on all but one data point and then tested on the single left out point	Using leave-one-out cross-validation to evaluate the performance of a machine learning model for predicting heart disease
Lemmatization	The process of reducing a word to its base form.	Lemmatizing patient notes to reduce each word to its base form (e.g., "running" becomes "run") for easier analysis.
Machine learning	A subset of artificial intelligence that involves training algorithms on data to enable them to make predictions or take actions without being explicitly programmed.	Using a random forest machine learning model to predict which patients are at risk for developing heart disease
MLP	Multi-Layer Perceptron. The unit of an artificial neural network is the perceptron which represent the single artificial neuron. MLP is comprised of multiple layers of fully interconnected perceptrons (with typically non-linear activation function) which also utilize the concept of backpropagation, collectively distinguishing them from a simple linear perceptron. The MLP also includes 3 layers: Input, hidden and output layers.	An MLP was used in a clinical medicine study to build a model that predicts the presence or absence of early sepsis in a burn's population.
Model	Refers to the result product created by an ML algorithm from its fed training data. This model maintains the mathematical relationships of the training data and can now make predictions on future unforeseen data	A sepsis model was generated that was able to predict the presence or absence of early sepsis in a burn's population with a high degree of accuracy and precision
Multicollinearity	A phenomenon in which multiple features (i.e., Independent /predictor variables) in a supervised model are highly correlated to each other, which can affect the model's accuracy and interpretability by overemphasizing certain features that represent the same or nearly the same thing	The model included two very closely related features (i.e., hemoglobin and hematocrit) with very high correlation (99%). Given that the model only had 3 feature inputs (hemoglobin, hematocrit, and blood pressure) for predicting heart disease, the extra emphasis on the two highly correlated variables (hemoglobin and hematocrit) may dilute the true contribution of the blood pressure variable and ultimately deteriorate the model's final performance
Multi-Layer Perceptron	See MLP	See MLP
Naïve Bayes	A machine learning classifier algorithm that makes predictions based on simple naïve assumption that the features are independent of each other (which is not likely the case in many studies).	A Naïve Bayes classifier may sometimes yield good results (depending on what data is involved). Additionally, it is important to note that as opposed to many of the other supervised algorithms (random forest, neural network, logistic regression, KNN, SVM, etc.) which have hyperparameters to tune, the Naïve Bayes algorithms are devoid of hyperparameters which further highlights their true simplistic approach
Natural Language Processing (NLP)	A subfield of artificial intelligence that involves using algorithms and other computational techniques to process and analyze natural language data, such as text and speech.	ChatGPT is a very powerful NLP-based chat platform that is scratching the surface of Artificial General Intelligence. Other NLP-based platforms include Siri, Alexa, etc.
Negative Predictive Value (NPV)	In essence describes the likelihood of no disease given a negative test. NPV along with PPV, F1 and Accuracy are prevalence-dependent statistical metrics NPV = (True negatives) / (True negatives + False negatives)	A negative D-dimer test was found to have a sensitivity of 100%, specificity of 8.8% and NPV of 100% for exclusion of pulmonary embolism in the proper setting. Hence, a negative result in such settings in essence rules out the disease.
Neural Network	Collectively these machine learning algorithms attempt to simulate the structure and function of the human brain. They are typically comprised of various artificial neuronal layers which include an input layer, intervening hidden layer(s), and output layer(s)	A convolutional neural network trained on colon cancer and normal colon cases is now able to identify new colon cancer cases (based on histological image features)
Normalization	A type of scaling used in ML. The typical normalization scaling approach (e.g. Min-Max Scaler) changes the range of values of the variable which may or may not change the shape of the distribution of the variable through (x – min) / (max – min) approach (e.g., changing the minimum value of the variable to 0 and its maximum to 1).	In medical research, normalization might be used to scale the units of measurements for different predictor/feature variables (that may have different scales) and scale them all to values between 0 and 1. e.g., some features may have values in the 100 s and some with values below 10, such as blood pressure values and creatinine values, respectively)
Object Detection	A type of computer vision task that involves identifying and localizing specific objects in images or videos.	In laboratory medicine, object detection could be used to automatically identify and locate different white blood cells in medical images, such as in a peripheral blood smear. These objects are many times seen as boxed entities within the single image.

(*continued on next page*)

Table 1 (*continued*)

Term	Definition	Example from medicine
Optimizers	An ML method that is used to adjust the parameters of a model to minimize a loss function. In essence it is used to improve its performance.	Examples of optimizers in ML include but are not limited to Stochastic Gradient Descent (SGD) and Adam
Overfitting	A phenomenon in which a machine learning model performs well on the training data but poorly on new or unseen data. This can occur when the model is too complex or when it has been trained on a limited dataset. Note: overfitting typically means not generalizable	An example is a model that performed well on the initial validation test (e.g., 97% accuracy and F1 of 91) while doing poorly on the follow-up secondary and tertiary generalization test sets (e.g., 80% accuracy and F1 of 73).
Parametric algorithm	An algorithm that makes assumptions about the underlying data distribution to make predictions or classify items. Parametric algorithms are often faster and more interpretable than non-parametric algorithms, but they may be less accurate if the data does not conform to the assumptions.	In medical diagnosis, a parametric algorithm (e.g., logistic regression) might be used to predict a patient's likelihood of developing a certain condition based on their age and other variables, if the data follows a normal distribution.
PCA (Principal Component Analysis)	PCA is a dimensionality reduction technique (an unsupervised learning method) that is used to transform and/or reduce the number of features in a dataset while retaining as much of the original information as possible	In an imaging study, PCA could be used to reduce the number of pixels in a large medical image, making it easier to process and analyze
Pixel	This is short for "Picture Element" and is the smallest unit of a digital image that can be independently assigned a color or intensity value. In other words, it is the smallest unit of information that can be displayed on a screen	In digital pathology space, the histologic image of a tissue sample taken from a patient with cancer could show individual cancer cells that are represented by pixels that are a different color or intensity than the surrounding healthy cells
Pooling Layer	In the context of convolutional neural networks (CNNs), a pooling layer is a type of hidden layer that is used to down-sample the spatial dimensions of the input data, reducing the number of parameters and improving the network's computational efficiency while preserving important information	For example, in VGG (a deep convolutional neural network), the pooling layers are used after every two or three convolutional layers to reduce the size of the feature maps and the number of parameters in the model
Positive Predictive Value (PPV)	This is also known as precision. In medicine, this can be seen as basically the likelihood of disease given a positive test. PPV along with NPV, F1 and Accuracy are classic prevalence-dependent statistical metrics. PPV = (True Positives) / (True Positives + False Positives)	For example, let's say we have a trained cancer prediction model that is now tested on another dataset of 500 images (mix of positive and negative cases). Out of those 500 images, the model predicted 150 images as positive for cancer. From those, 120 images were shown to be truly positive (cancer was present) and 30 images were false positives (no cancer was present). Therefore, the PPV of this algorithm is 120 / (120 + 30) = 120/150 = 0.8 or 80%
Precision	Another name for Positive Predictive Value (PPV)	See Positive Predictive Value (PPV) above
R^2	In the context of regression analysis, the R^2 metric is a measure of the goodness of fit for a linear regression model, indicating how well the model fits the data. An R^2 value of 1 indicates a perfect fit, while an R^2 value of 0 indicates no relationship between the predictor and response variables.	For example, the R^2 can be calculated within a regression model to see how much of the weight variation is explained by height. This was shown to be 0.8 which shows that these two variables (height and weight) are relatively highly correlated in this example.
Random Forest	An ensemble tree-based machine learning algorithm that combines the predictions of multiple randomly generated decision trees, which can potentially improve the model's accuracy and reduce overfitting.	A random forest model was found to be one of the best models for predicting early sepsis in our burn's population.
Regression	In the context of statistics, regression is a type of analysis that is used to model the relationship between one or more predictor variables and a response variable. However, in machine learning this refers to a model that is used to predict a continuous numerical outcome/target (rather than discrete classes).	The classic regression model example that peoples use is prediction of a home price based on the house's zip code, number of bedrooms and bathrooms and square footage. In lab medicine, this same concept can be used to build a random forest regression model that can predict the cost of a laboratory test (given certain other variables).
Regularization	A technique used in machine learning to minimize overfitting. Ultimately, this can be thought of to address the bias-variance tradeoff issue by adding a penalty to the loss function of a model to reduce its variance at a sacrifice of a small amount of bias.	For example, this is accomplished by adding a penalty term to the model's loss function which ultimately leads to a more generalizable model (less likely to overfit). Common examples of this includes LASSO, Ridge, and Elastic Net.[1]
ROC/AUC	A ROC curve is plotted by assigning the model's various sensitivities and 1 minus specificities (i.e., False Positive Rates) at different thresholds (i.e., cut-off values). After which the area under this ROC curve (ROC-AUC) is calculated and used as a metric to evaluate the classification model's global performance potential (reflecting on how the overall model can separate the output classes).	Two delayed graft function ML models were compared based on their ROC-AUC and one was shown to be 0.56 while the other showed an ROC-AUC of 0.72. The 0.56 is close to predicting by random chance (i.e., 0.5) so the 0.72 model is found to be a better performing model in this case. Note: the closer the ROC-AUC to 1, the better.
Root-mean-square error (RMSE)	A metric that measures the average difference between the predicted values and the true values in a dataset when using a regression model. The lower the RMSE value, the better (which signifies a better fit)	In a regression model that predicts the systolic blood pressure of a patient, the root mean square error could be used to measure the model's performance. Note: they are generally not used for comparison of models with different targets (in those cases R^2 comparison is generally preferred)
Sample	A representative portion of a population, selected for analysis to draw conclusions about the population as a whole	A sample population of all septic patients were taken to be used for building a sepsis prediction ML model.
Scaling	See the terms Normalization and Standardization for more details	e.g., Normalization and Standardization
Secondary/Tertiary Test Sets	In machine learning, a set of data that is used to evaluate the performance of a trained model. The secondary or tertiary testing sets (i.e., generalization test set) are separate from the initial training/validation test set. These are used to assess the model's ability to generalize to new data.	Sample of blood from a cohort of patients was used to build an ML model whose subset of initial validation test was shown to identify positive cases. This ML model was then subjected to additional tests (secondary and tertiary tests) from other sample populations of patients.
Sensitivity		

(*continued on next page*)

Table 1 (*continued*)

Term	Definition	Example from medicine
	The ability of a test to accurately identify individuals who have a specific condition or disease. In other words, "likelihood of positive test given disease". This is also known as positive recall or True Positive Rate (TPR) in the ML world. Sensitivity = = (True Positives) / (True Positives + False Negative)	In the following example, this cancer prediction model has resulted in 88 True positive cases. These are positive cases in real life that the model is predicting correctly as positive. Sensitivity = 88 / (88 + 8) = 92% Predicted Class + - True + True Positive: 88 False Negative: 8 Class - False Positive: 14 True Negative: 112
Sentiment Analysis	A natural language processing (NLP) approach that helps identify and extract subjective information from text, such as the emotional state or the overall attitude towards a topic. This could be an expression, that is, the opinions, appraisals, emotions, or attitudes towards something.	A sentiment analysis approach was used to acquire patient feedbacks on a new drug which is helping us determine the overall patient satisfaction with this drug and to also identify potential side effects
Specificity	The ability of a test to accurately identify individuals who do not have a specific condition or disease. In other words, "likelihood of negative test given no disease". This is also sometimes known as True Negative Rate (TPR) in the ML world. Specificity = (True negatives) / (True negatives + False positives)	In the following example, this cancer prediction model has resulted in 112 True negative cases. These are negative cases in real life that the model is predicting correctly as negative. Specificity = 112 / (112 + 14) = 89% Predicted Class + - True + True Positive: 88 False Negative: 8 Class - False Positive: 14 True Negative: 112
Standardization	A scaling preprocessing approach. Standard Scaler is a commonly used standardization approach that transforms the variables through the (x – mean) / standard deviation approach yielding a mean of 0 and a standard deviation of 1 without distorting differences in their ranges	In medical research, standardization might be used to scale the units of measurements for different predictor/feature variables (that may have different scales) and scale them all to a mean of 0 with a standard deviation of 1 e.g., some features may have values in the 100 s and some features have values below 10, such as blood pressure values and creatinine values, respectively). These values are all scaled into a mean of 0 with a standard deviation of 1 before feeding them into machine learning
Stemming	A preprocessing natural language processing (NLP) technique for text data that removes common prefixes and suffixes from words (reducing words to their base form; "stem").	An example in the pathology world would be reducing words such as "cancer," "cancerous," "carcinoma" to "cancer" as the root word
Stop Words	Common, uninformative words that may be excluded from natural language processing (NLP) tasks to improve the efficiency and the accuracy of the end model. Words such as "a," "an," "the," "in," "on," are of little value (do not contribute significant meaning or context to the text) and often removed in the preprocessing step.	In a pathology search tool, removing words such as "study", "research", or "report," may not carry much meaning in terms of the specific case for a pathology report (e.g., looking for "sarcoma cases")
Supervised Learning	Supervised learning as the name implies uses data that makes use of "labeled" targets (human designated).	Examples include Classification and Regression ML Tasks
Support-Vector Machine (SVM)	By defining an optimized hyperplane this supervised approach enables us to ultimately differentiate the classes of interest (i.e., a classification task) by optimizing the margin that separates these classes.	An SVM model using an RBF kernel was found to be one of our best performing sepsis prediction models.
Synthetic Data	New data that is generated by a machine, rather than being collected from the real world. The machine first learns the collective mathematical relationships of the real data which is then used as a function (i.e., model) to create new data that maintains the real data's collective characteristics. Synthetic data can be used to supplement or replace real-world data in certain machine learning tasks.	The use of such validated data may be very useful for drastically expediting pilot studies and can revolutionize our current approach in certain research domains (since such data can be more readily accessible to all, considering its nature of not representing any individual patient info)
Target	In ML studies, this refers to the dependent variable (AKA "outcome" measure) within a supervised model	In a dataset that is building an ML model to predict "sepsis" based on patient's hemoglobin, temp and blood pressure, the "sepsis" variable (i.e., positive, or negative cases) is the Target variable
Tensor	A mathematical object that represents a multidimensional array of numbers. Tensors are often used in neural network ML algorithms to represent high-dimensional complex data such as images or texts (e.g., through word embedding)	In pathology whole slide imaging studies, tensors can be used to represent the histopathology images (represented as multi-dimensional arrays, with the dimensions representing the height, width, depth, and other characteristics of the image)
Testing Set	See the term "Secondary/Tertiary Test Sets" above	See the term "Secondary/Tertiary Test Sets" above
Threshold	This is the value (e.g., probability) that is used by the ML model to make predictions. For example, in most binary classifiers, the model outputs a probability for each class as being called positive (1) or negative (0). By default, most ML model probability thresholds are set to 0.5 which means that if the probability of a class is found to be >0.5, that input is	Investigators who initially had created a hypothetical Ebola prediction model (0.5 threshold) found it to be 98% sensitive and 94% specific. Since they did not want the model to miss any real positive cases, the new goal was to improve its sensitivity to 100%. To do this, they adjusted the model's threshold from 0.5 to 0.4 which made the model 100% sensitive (decreased the false negative cases) and 93% specific.

(*continued on next page*)

Table 1 (*continued*)

Term	Definition	Example from medicine
	then classified as the positive (i.e., 1) class while those below that threshold will be categorized as negative (i.e., 0) classes.	
Tokenization	This is a routine preprocessing task that is applied to NLP tasks to breaks down the input texts into individual words or symbols. Tokenization can also be applied to images. In a pathology search tool, tokenization could be used to extract specific information from the pathology reports (e.g., diagnostic codes, or descriptions of specific histologic findings). The term tokenization can also refer to replacing patient's sensitive data with unique identifiers (goal being to retain the essential information while protecting patient privacy)	For example, a pathologist is using a search tool that incorporates NLP. Their goal is to search for all cases of "basal cell carcinoma" within their hospital system. The tool would then tokenize the keywords above which will then enable it to search through the database for any images or reports that contain those specific words
Training set	A subset of the data used to train a machine learning model. See "Train-Test Split" for more details	In colon cancer image study, 500 images of colon cancer and 500 cases of normal colon were used in an 80–20 train-test approach (i.e., 80% used for training and 20% used for initial validation test) to train a deep neural network that will create the model that is able to distinguish these two classes (cancer and normal).
Train-Test Split	Supervised ML studies follow the process of train-test split which means that a subset of the data is used to train the model and then the remaining is used to do the initial validation testing on the model	For example, the 80–20 split, in which 80% of the Cancer images and normal tissue images are used to train the model and the left over 20% is used to do the initial validation test to assess the model's performance
Transfer Learning	An ML method (commonly used in many deep learning image studies) in which a pretrained ML model (trained and optimized on one task) is used as the starting point to retrain it for a different task	The ResNet-50 neural network model (which was trained on the ImageNet data to distinguish various images of animals, vehicles, buildings, etc.) is now retrained through transfer learning from histological images of colon cancer and normal colon cases to make a new colon cancer predictor ML model
Unsupervised learning	Machine learning algorithms/methods which are used to identify patterns on "Unlabeled" data output variables (i.e., no human input)	Using the k-means clustering method the computer can identify different subtypes (i.e., clusters) of a particular disease
Validation test	This refers to the initial testing of a trained ML model to determine how well the model initially performs on unseen data See "Train-Test Split" for more details	For example, the 80–20 split, in which 80% of the Cancer images and normal tissue images are used to train the model and the left over 20% is used to do the initial validation test to assess the model's performance. Note: especially for small to intermediate sized data the initial validation test set is not generally enough and additional test sets (e.g., secondary & tertiary test sets) are required to truly assess the model's generalizability.
Variance	In statistics, this refers to a measure of spread of a data's distribution by quantifying how much the individual values in a dataset deviate or vary from the mean or average value In ML terminologies, this refers to the degree to which the model's predictions vary based on changes in the input data or the model's parameters and it is associated with model complexity as the model is trained (i.e., increased training typically leads to a more complex model which translates to increased variance). See Bias-Variance Tradeoff for more details	A cancer prediction ML model was found to be not generalizable (i.e., overfitted) which is likely due to overtraining of the model and becoming markedly complex (i.e., increased variance led to the overfitting in this case). See Bias-Variance Tradeoff for more details

Data preprocessing

Data pre-processing is a crucial step in the machine learning life cycle. One of the great challenges within the field of healthcare is access to a clean and complete dataset. The quality of the data is extremely important as this can influence the ability of the model to learn and ultimately become generalizable.

The type of preprocessing is also dependent on the data type and their respective employed ML algorithm. Common pre-processing tasks within image classification may include image resizing, cropping, normalization, image augmentation (e.g., image rotation, etc.), noise reduction (e.g., blur), and color space conversions while those for text data NLP classification tasks may need tokenization, lowercasing, stop words removal, lemmatization, stemming, and encoding (i.e., converting to numbers through one-hot encoding or word embeddings). On other hand, the common pre-processing tasks in tabular data include tasks that may concentrate on minimizing noise within the data by finding the optimized number of features (i.e., by employing certain feature selectors or feature transformation tools), finding ways to best deal with missing values (i.e., removal versus imputation) and outlier data, along with important scaling tasks (e.g., normalizing the data values through a scaling transformation). Below we will dive deeper into some of these preprocessing approaches (see Table 2).

2a. Image Data Preprocessing: Image preprocessing techniques such as resizing and cropping play a crucial role in the development of an optimized machine learning model. Resizing images to a uniform dimension ensures that the neural network receives consistent inputs. Cropping, on the other hand, allows for a more focused analysis, zeroing in on a specific region of interest within an image. These techniques can help to reduce noise and irrelevant information in the data, therefore improving the accuracy of the model. Additionally, other data augmentation techniques, such as flipping, rotating, and normalizing can also be applied to increase the diversity of the training data, leading to a more robust model that can better generalize to unseen data. The implementation of these preprocessing techniques is vital for the successful deployment of an image-based machine learning model.[13] Besides the aforementioned tasks, it is also very important to address any bias within the image selection process or any image quality issues that may need to be addressed as noted by Wright et al. and others.[14]

2b. Text Data Preprocessing: For NLP tasks (i.e. text data) some of these preprocessing tasks can be grouped into low and high-level tasks[4] and it is the combination and interplay between these two essential groups of tasks that ultimately will best optimize the NLP model. Low-level tasks are those that involve the manipulation of individual words or smaller units of language which often involve the processing of raw text data and include tasks such as

Table 2
Common preprocessing tasks for tabular/numerical datasets.

Type of problem	Solution / Potential preprocessing tasks to address this	Special points
Missing values	Remove the missing values[20]	Generally preferred if the data size is not drastically compromised (when data is removed). If the data size is compromised, may consider other approaches such as imputation
Missing values	Impute the missing values (assign a value by inference) [21]	Typically, should not be done on secondary / generalization dataset since it may skew or exaggerate the performance of the ML model
Different variables having different ranges of their numerical data	Standardization scaling through Standard Scaler approach or Normalization scaling through the Min-Max Scaler approach for the data values	There is difference between Standardization and Normalization. Standardization scaling (e.g. Standard Scaler) transforms the variables through the (x – mean) / standard deviation approach yielding a mean of 0 and a standard deviation of 1 without distorting differences in their ranges. In contrast, the typical normalization scaling approach (e.g. Min-Max Scaler) changes the range of values of the variable which may or may not change the shape of the distribution of the variable through (x – min) / (max – min) approach (e.g., changing the minimum value of the variable to 0 and its maximum to 1).
Text that needs to convert to a numerical value	Binarization	This is a great option for binary target columns (e.g., Cancer vs No Cancer) or certain binary feature columns (Shortness of Breath being present or absent, etc.). For example, a single column of Cancer cases which contains "Positive" and "Negative" cases can be converted to a binary column of 0 s (Negatives) and 1 s (Positives), as such: Cancer | The single "Cancer" Column to the left is converted to a Binary column of 0s and 1s on the right | Cancer Positive | 1 Positive | 1 Positive | 1 Negative | 0 Negative | 0
Text that needs to convert to a numerical value	Encoding	This could sometimes become problematic since the true representation of each value may not be presented in the most accurate way (especially since some of these may be seen as having higher significance in their measured values for the algorithms that could ultimately deteriorate the final ML model). For example, a single column of Colors which contains "blue", "green" and "brown" values is converted to 3 separate numerical values of 1, 2 and 3 (ultimately incorrectly signifying that encoded values listed below for brown (3) is 3 times as significant as blue (1), etc., for certain algorithms (e.g., k-NN). Colors | The single "Colors" Column to the left is converted to the following encoded column | Colors blue | 1 blue | 1 green | 2 Brown | 3 green | 2
Text that needs to convert to a numerical value	One-Hot Encoding: Creates a new binary column for each unique category in the data [22]	One-Hot encoding can address the simple encoding issue (shown above) by separating the individual characteristics within the column into their own unique binary individual columns. For example, a single column of Colors which contains "blue", "green" and "brown" values is converted to 3 separate binary columns titled "Colors_blue", "Colors_green", and "Colors_brown" as such: This now maintains the true intrinsic values of the assigned colors which ultimately translates to a higher likelihood of better representation of these values within the ML model. This approach is not without its own issues as it can add many additional (sometimes unnecessary) columns to the dataset and become computationally very demanding. Colors | The single "Color" Column to the left is converted to the following 3 separate binary columns on the right | Colors_blue | Colors_green | Colors_brown blue | 1 | 0 | 0 blue | 1 | 0 | 0 green | 0 | 1 | 0 brown | 0 | 0 | 1 green | 0 | 1 | 0
Multicollinearity	Multicollinearity assessment tool	Multicollinearity in short, refers to redundant independent variables /features within a dataset.[19] High correlations between certain features may also sometimes deteriorate an ML model's performance as certain feature's contributions may overpower others (especially when the number of features/independent variables are low). For example, in a hypothetical kidney disease dataset the following 3 features (Hemoglobin, Hematocrit and creatinine) are present and hemoglobin and hematocrit (not surprisingly) show a 99% correlation. If both Hemoglobin and Hematocrit are kept in the final model, their summative feature contribution to the target (i.e., kidney disease) may dilute the true contribution of creatinine to the target. By removing either the Hemoglobin or Hematocrit value from the final dataset, the multicollinearity issue noticed here will be addressed and the final model may potentially become more predictive.
Features that are not contributing much to the overall data (i.e., noise)	Feature selection tools	By employing certain feature assessment tools (e.g., ANOVA F-value-based Select Percentile or the Random Forest Feature Importance's feature selection processes), we can identify which set of features are most contributing to a given target which will help to potentially minimize the noise (the unnecessary features) in the dataset.

tokenization. Tokenization refers to splitting text into individual words or smaller units that enable the machine to then treat each instance as a separate value, analogous to how a machine may use an array of numbers in tabular or image classification tasks. Another important low-level task is the removal of stop words which are words that generally do not add a great deal of value[15] (Fig. 1). Stop words are words that are commonly used in a language but do not carry much meaning on their own (such as "a", "an", "the", "on", "in", "at", "and", "or", "but", etc.). Other low-level NLP tasks include stemming (reducing words to their base form), and part-of-speech tagging (assigning a grammatical role to each word in a sentence) while the high-level tasks involve the interpretation of the overall

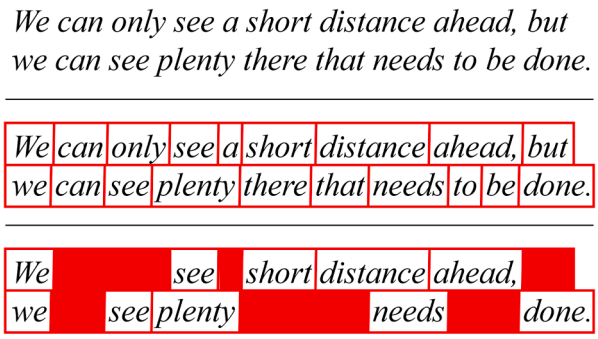


Fig. 1. The same quote is shown three times: the top is the sentence as-is, the middle shows the sentence after it has been tokenized and the bottom shows the stop words removed.

meaning of a piece of text or the relationship between words and concepts which employ tasks such as language translation, text summarization, and sentiment analysis.

2c. Audio data preprocessing: Several steps are typically employed to ensure that audio data is suitable for use. Most will start by removing any background noise from the audio recordings, which can be achieved using techniques such as noise reduction or filtering. Another important step is to segment the audio into meaningful chunks, such as individual words or phrases. Feature extraction on the audio data is also performed, such as calculating the mel-frequency cepstral coefficients (MFCCs), to extract relevant information and represent the audio data in a format that is suitable for machine learning.[16]

2d. Tabular/numerical data preprocessing: The cleaning tasks (i.e., preprocessing) of the numerical data include but are not limited to data noise reduction (e.g., through feature selector tools, etc.), missing data replacement (i.e., imputation) or removal, scaling tasks, text to number conversions, and statistical assessment of the features (e.g., multicollinearity minimization). Collectively these approaches can help optimize the ML training process for tabular studies and ultimately lead to a more generalizable model (Table 2). Many of these have traditionally been confined to various manual (non-automated) approaches but now with the advent of certain powerful data science applications[17], we are finally able to expedite many of these tedious processes through automated standardized validated approaches (Figs. 2 and 3). These applications can provide us with a very user-friendly and scientifically sound tool kit that can readily address the preprocessing tasks noted above. These tools ultimately enable the user to expedite their data cleaning phase within their respective study which not only translates to a faster deployment process but more importantly to a higher likelihood of achieving a more generalizable machine learning model.

As noted earlier, these preprocessing tools are very data-type specific which also translates to their corresponding compatible ML algorithms. A better understanding of the inner workings of these ML algorithms and their capabilities and limitations is a key step in the better use of these tools. However, before diving deep into these algorithms, it is also essential to understand the different ML categories that they belong to since that will give us a better big-picture view of how these powerful tools operate.

Machine learning categories

Machine learning is typically categorized into the following three groups: 1. Supervised Learning, 2. Unsupervised Learning and 3. Reinforcement learning. Supervised learning as the name implies uses data that makes use of labeled targets (human-designated). These can further be subcategorized into classification (i.e. output is a discrete class such as cancer versus no cancer) and regression[21] (i.e. a continuous numerical output such as the predicted cost of a given lab test) [22]. In contrast to supervised learning, unsupervised learning does not involve any labeled output data (i.e., unlabeled with no or less human input) and is usually used to identify potentially unrecognized clusters, patterns, outliers, or feature subsets. The most common unsupervised methods employed within ML include clustering (e.g., K-means; Fig. 4), dimensionality reduction (e.g., Principal Component Analysis; PCA), and outlier detection which can sometimes also help identify optimized feature sets. These unsupervised ML tasks along with others can be employed to optimize the data as it feeds into the supervised algorithms which can help enhance their final performance measures. The third category is reinforcement learning[23] which utilizes a trial-and-error process and employs a "reward" parameter to continuously evaluate the relationship between its input and output values and decide the best way to optimize itself[24]. Given that reinforcement learning is not currently much used within medicine, we will mainly focus on the supervised methods within this article. For more details on unsupervised and reinforcement learning, please refer to the glossary of terms (Table 1) along with some excellent other cited review articles that further focus on those approaches[31].

It is also important to reemphasize that the choice of the supervised algorithm used for a given study will depend on multiple factors which include but are not limited to the datatype (text, numerical, or image), the data quantity, and ultimately its intended use.

Supervised machine learning

The concept of a train-test split is fundamental to supervised machine learning studies. Train-test split (e.g., 80–20 = 80% training and 20% testing) refers to how a portion of the data is assigned to the training arm of the algorithm (to help map the features to the target of interest) while the leftover test data is then used to validate the performance of the trained model (known as validation testing in the ML world). For more detail on how this works along with other key concepts such as cross-validation and bias-variance tradeoff please refer to the glossary of terms (Table 1) and other excellent review articles (cited below) on this topic.

As was noted, supervised learning can be divided into two sub-categories, classification, and regression[1]. It is important to note that some ML algorithms are only capable of performing classification tasks (e.g., Logistic Regression), some are regression task-specific (e.g., Linear Regression) and some can do both (e.g., k-nearest neighbor, Random Forest, or Neural Networks). It is also important to note that there are key differences between the statistical performance measures employed in a classification task versus those used in regression modeling. The statistical performance of a model in a classification study is usually confined to a confusion matrix-based approach. This means that the true positive, true negative, false positive, and false negative measures from the confusion matrix acquired from the ML model are then used to calculate various known statistical metrics (e.g., accuracy, F1, sensitivity, specificity, Matthew's correlation coefficient, etc.). In contrast, regression models are assessed through other performance metrics such as R^2, Mean Squared Error, etc.

It is also of critical importance that AI/ML applications include accurate, precise, complete, and generalizable information[25,26]. Following the concept of "garbage in, garbage out," if incorrect or poor-quality data is used the output will likely be a suboptimal ML model that will not likely address our study need. Hence, as was noted earlier, the use of data preprocessing tasks is also essential to minimize such issues [27].

Another key concept around supervised learning is how to minimize the overfitting phenomena (i.e., maximize the generalizability of the ML model) [25,26]. This can be achieved through various tasks which include cross-validation along with the use of additional testing datasets (e.g., secondary and tertiary test data[28] to assess generalizability, especially critical for smaller-intermediate sized datasets) [25]. Another important

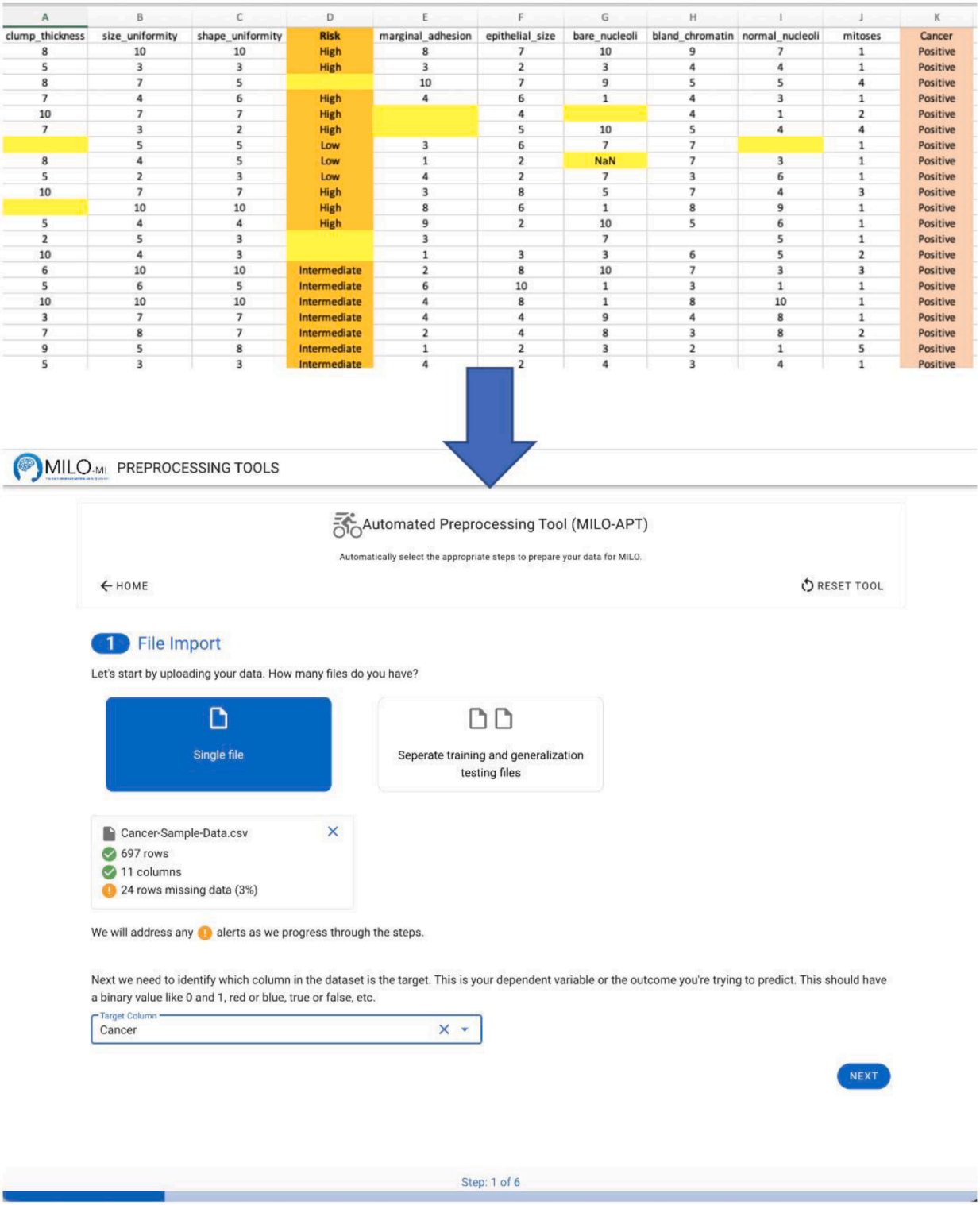


Fig. 2. Note the missing values (including NaNs; highlighted in yellow) and the non-numerical data (highlighted in orange) within this dataset that is uploaded into this Automated Preprocessing Tool. Once uploaded, step 1 shows certain details about the dataset (e.g., the number of rows and columns along with the percentage of missing values).

future approach to enhance and maintain generalizability would include the concept of pooling multiple sites in the training process with continued assessments of the models (i.e., a federated ML process) as highlighted by Yang et al.[29] Lastly, once a generalizable model is deployed, it is critical to also constantly assess its performance since data may change over time (especially for certain tabular data) and lead to a deteriorating ML model (a concept known as data drift) [30]. This concept is also not uniform throughout the ML models and very much can be data-type dependent (i.e., tabular data will likely have a higher tendency of drifting while this phenomenon may be less common in certain relatively standardized data such as images). Since in many instances, the data type governs the ML approach (e.g., the gold standard approach

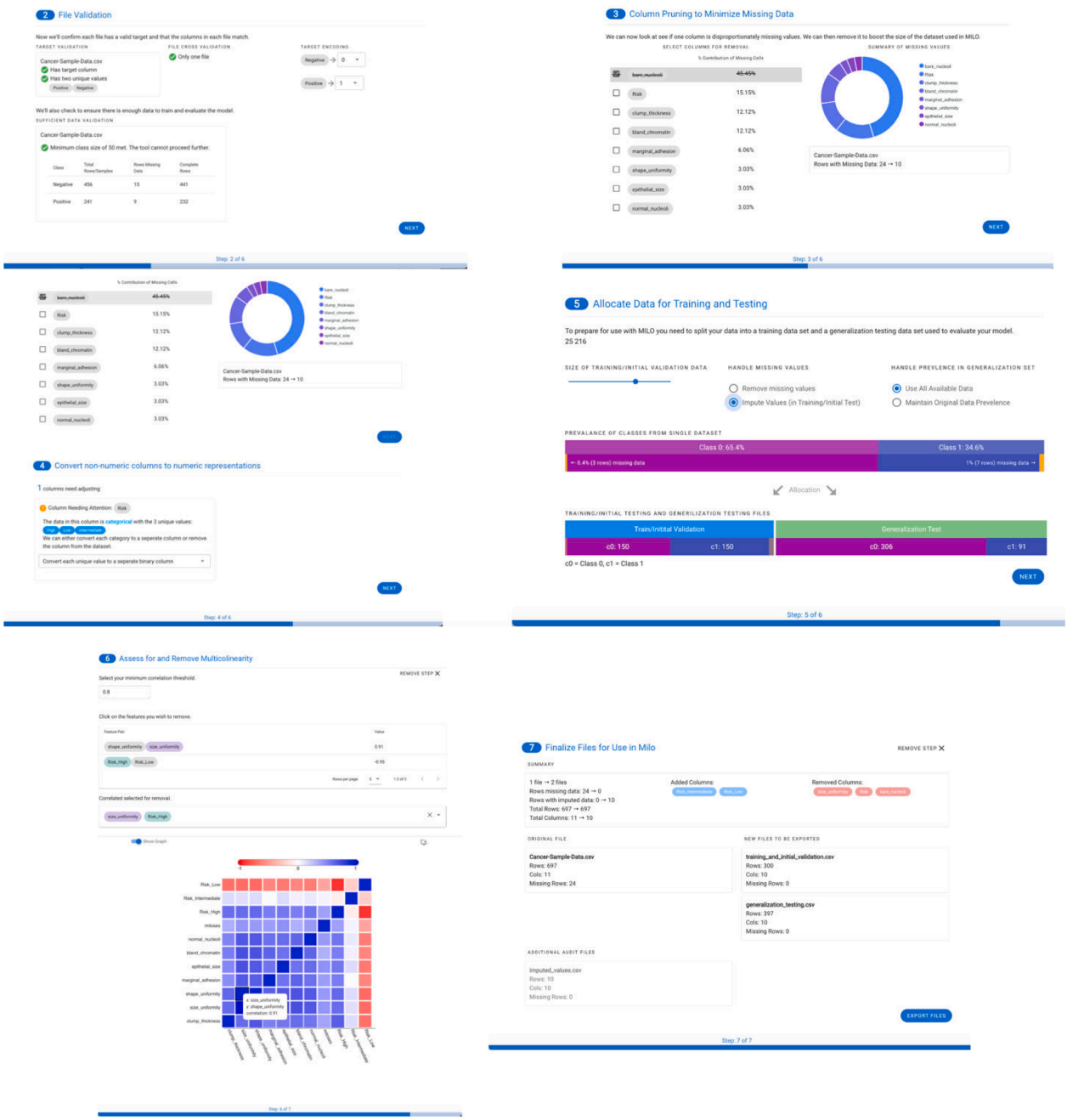


Fig. 3. In step 2, the non-numerical target column (i.e., "Cancer") is changed to 0 s (negatives) and 1 s (positives) through a simple binarization process. In step 3, each of the individual features and their contributions to the missing values are noted. In this case the feature with the most missing values was marked for removal ("bare Nucleoli" column which was contributing 45% of the missing values) to minimize the number of values that we will later impute. In step 4, the "Risk" column which contains text values (Low, intermediate, and high) are converted to the appropriate numerical values in separate columns through the one-hot-encoding process. In step 5, the single dataset uploaded is split into two separate datasets (one for training/initial validation testing and the second one for secondary/generalization testing) to prepare it for the later machine learning model building and testing phases that will follow. Additionally, the missing values are also automatically imputed in this case rather than removed (orange bar) through the embedded multiple-iteration imputation tool. In step 6, multicollinearity is assessed and the following two features with high multicollinearity ("size uniformity" & "high risk") are marked for removal in the final datasets created (shown in the summary in step 7).

for most image studies is convolutional neural network), a better understanding of the various supervised ML algorithms is also key as we embark into this space.

Supervised machine learning algorithms

Common supervised machine learning algorithms (specifically linear regression, logistic regression, neural networks, K-NN, support vector machine, gradient boosting machine, and random forest some of which are depicted in Fig. 5) are further described below:

Neural networks

Collectively these machine-learning algorithms attempt to simulate the structure and function of the human brain. They are typically comprised of various artificial neuronal layers which include an input

Fig. 4. The figure depicts the iterations of a k-means clustering algorithm used in unsupervised ML.

layer, intervening hidden layer(s), and output layer(s). These can be used for a variety of tasks and can be applied to all data types (i.e., image/video, text, audio, and tabular data). Most common neural networks employed within the ML arena include the Multilayer Perceptron and Convolutional Neural Networks.

Convolutional neural network (CNN)

A convolutional neural network (CNN) is specially optimized for recognizing data motifs and smaller patterns[31]. These include patterns in images, videos, time-series data, genomic data, or other types of data with grid-like characteristics[32]. Generally, CNNs are like other feed-forward neural networks but have an extra layer called a convolutional layer that aids in pattern recognition. A convolution is simply a mathematical operation performed on data fed into that specific detection layer and a kernel[33]. The kernel is a small matrix or window that is initialized to have certain numerical weights in each subframe of the window. This kernel acts as a feature detection filter as it moves over the input and processes that data to produce a resulting feature map through its activation functions. Next, a method called pooling is used to summarize data from the feature detection layer as well as to provide invariance to minute transformations in the input[34]. This method is typically performed in multiple stages interspaced with feature detection layers where the final stage reduces the dimensionality of the input down to a single feature vector. This vector is then used as the input for a dense neural network for various classification purposes [35,36]. However, these neural networks are not typically built from scratch for many of our pathology image classification tasks since preexisting optimized neural networks (such as ResNet-50, etc.) can be retrained through the transfer learning approach to expedite this process while providing very

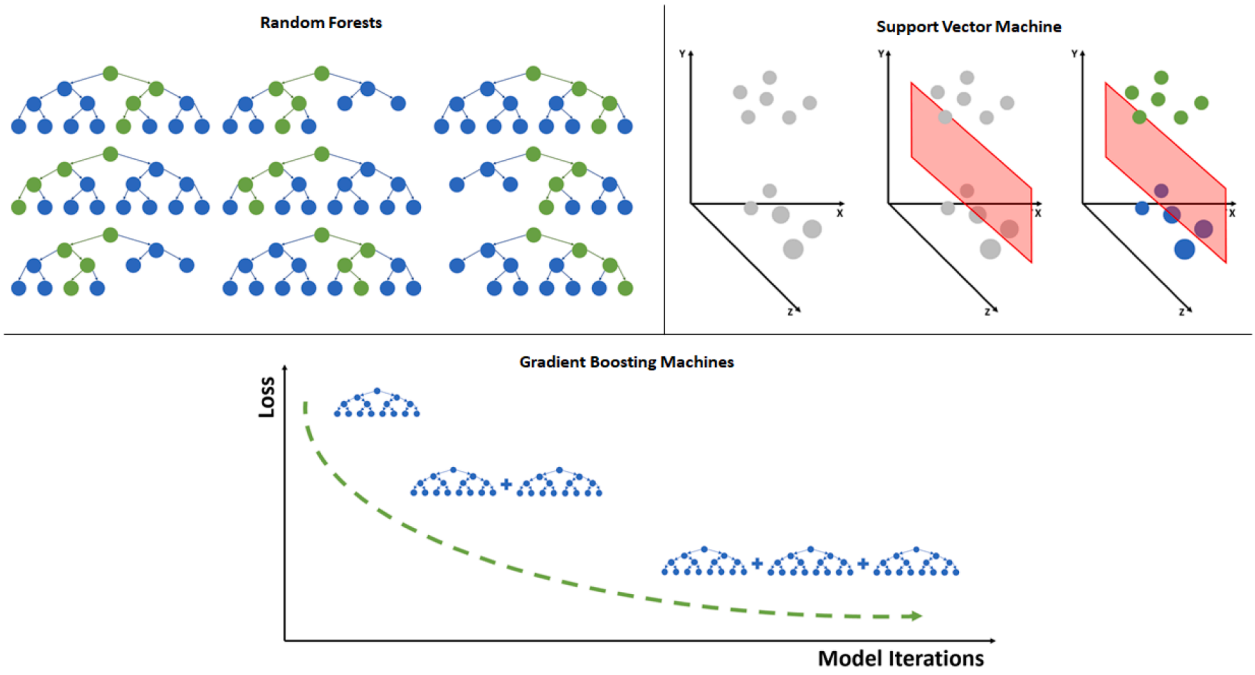


Fig. 5. Visual representation of some common supervised ML algorithms. The top left corner demonstrates the logic path of a random forest algorithm, top right corner shows a hyperplane separation of a support vector machine, and the bottom depicts a gradient boosting machine.

capable predictive tools within this space (for more details on transfer learning, see Table 1).

Neural networks have revolutionized our machine-learning capabilities within images, video, text (i.e., Natural Language Processing), time series, and certain other complex tasks but they are not always the best solution for all data types. This is especially true within the traditional tabular data space which is one of the most common data types in laboratory and clinical medicine. Although some neural network approaches (e.g., multilayer perceptron) are very promising within the tabular data space, several of the non-neural network algorithms have also been shown to be just as good and sometimes better. These include some of the more traditional methods such as K-nearest neighbor, logistic regression, linear regression, polynomial regression, naïve Bayes, and support vector machine along with some more advanced ensemble methods such as random forest and gradient boosting machine.

K-nearest neighbor (k-NN)

The k-nearest neighbor (k-NN) algorithm is a simple, yet powerful, clustering technique for classification and regression problems. k-NN finds uses in areas such as content retrieval, data mining, and binary or multiclass prediction[37]. In utilizing the k-NN algorithm, the k-parameter is a user-specified constant, which represents the number of nearest neighboring data points. Overall, the k-NN algorithm is highly flexible and can be used for a variety of tasks, including classification and regression tasks. Another benefit to k-NN is that the algorithm is non-parametric, meaning that it does not make any assumptions about the underlying data distribution[38]. It is also able to find non-linear relationships since the algorithm itself is non-linear in nature[39]. Unfortunately, k-NN is not without its disadvantages. The algorithm can be computationally expensive, particularly with large data sets. It is also sensitive to poorly selected k values, which may reduce their performance and decrease their accuracy. Finally, k-NN lacks recall, which requires the algorithm to reevaluate data it has seen before.

Linear regression

This technique is one of the oldest and simplest statistical approaches used in regression studies within supervised machine learning[25,40]. In its simplest form, the relationship between a single independent variable (the feature, e.g., height) and the dependent variable (the target, e.g., weight) is evaluated by finding the best-fitted straight line that minimizes the error between predicted and actual values (lowest error sum). Multiple linear regression is a variation of this that displays the relationship of a single outcome variable (i.e., target) to two or more independent variables. Certain regression metrics such as root mean square error and coefficient of determination (R^2) can then be used to evaluate the performance of these models [41]. R^2 describes the amount of variability in the outcome variable that can be explained by the independent variables. In short, how well the regression model fits the data.

These regression techniques are relatively simple and efficient in finding linear relationships but have their own shortcomings[40]. Mainly, it is not generally useful when relationships are nonlinear between features and targets[42]. Polynomial regression methods can help in such nonlinear data situations using their input variable exponentials (i.e., the independent variables are mapped to the target variable as nth degree polynomial).

Logistic regression

In contrast to linear regression which is used for continuous dependent variables (i.e. targets), logistic regression is confined to classification tasks (i.e. targets that are discrete classes) [43]. For a binary classification task, this method produces a probability score that can then be translated to a given binary output (e.g., 0 for negative for cancer and 1 for positive for cancer) through some default threshold (e.g., a 50% or more scores will be assigned to the 1 or positive cancer class while <50% will be placed into the 0 or negative class). The main objective is to find the best-fitting curve that can optimize the separation of the binary class of interest. This approach is not just confined to binary tasks and can also be applied to multiclass problems. However, to apply this approach to multiclass studies one needs to either employ the "One-vs-all" (also known as One-vs-Rest or OvR) approach or the "SoftMax regression" method (also known as multinomial logistic regression) [25, 44].

Support vector machines (SVM)

By defining an optimized hyperplane this approach enables us to ultimately differentiate the classes of interest (i.e., a classification task) by optimizing the margin that separates these classes. However, a variation of which (Support Vector Regression) can be applied to regression studies [45]. In nonlinear settings, this method can also employ kernels when the data is not easily separable by allowing the data to be transformed into another dimension which can enhance the dividing margins between these classes.

Random forest

This supervised ensemble tree-based approach can be applied to both classification and regression tasks. Multiple decision trees (hence the forest) are generated at random using features from a data set. Each decision tree makes its prediction or vote. When used for classification, the category with the most votes is chosen as the predicted outcome[46]. In the case of regression, numerical values are averaged together to make a prediction. Theoretically, the more trees there are in the forest, the more accurate the prediction. However, this could also lead to overfitted models that will deteriorate the generalizability [15].

Gradient boosting machine (GBM)

This is another ensemble tree-based supervised approach that can be used to solve both regression and classification problems[25, 47–50]. GBM sequentially creates new trees from an ensemble of weaker trees. Every time the ensemble adds a new tree, its model complexity increases, and its overall bias decreases. Theoretically, each new model fits more accurately with new observations which ultimately leads to enhanced overall accuracy. Algorithms based on this boosting principle are effective in reducing bias and variance that occur when applied to small or unbalanced datasets[51]. However, the limited number of tuning parameters may make them more prone to overfitting as compared to other approaches such as the random forest [52].

Summary and future direction

In summary, machine learning has the potential of revolutionizing the field of healthcare by enabling the analysis of large and complex datasets. Using different data types, preprocessing techniques, and machine learning algorithms, it is possible to gain valuable insights and make more accurate predictions. With the use of supervised and unsupervised learning methods, a wide range of tasks can be performed, including classification, clustering, and regression. However, in addition to embracing well-established tools, it is also important to continuously embrace new tools and concepts within these arenas which include but are not limited to reinforcement learning, Automated machine learning (Auto-ML), and the synthetic data realm. These exciting new areas (Auto-ML, reinforcement learning, and synthetic data creation and usage) of data science are starting to complement the traditional approaches and will help in expediting the adoption process of the various machine learning tasks within our healthcare space.

However, healthcare poses unique challenges. One of which is data drift, which is when input data changes over time, eventually becoming

significantly different from the model's original training and validation data. This happens for a variety of reasons but regardless of the cause, it is crucial for us to be aware of this phenomenon and to employ techniques to identify data drift and mitigate its effects. Such techniques include active or continuous learning approaches which allow the model to continue to adapt and improve over time. Within the clinical ML arena, the ML Operations (ML-Ops) arm can help address some of these needs as more and more of these advanced analytic tools are being deployed. The ML-Ops platforms not only address automation and deployment streamlining needs, but also the monitoring and maintenance of the deployed machine learning models.

Overall, the combined use of these new tools along with our traditional machine learning approaches has the potential of addressing many of these challenges and greatly improving patient care, paving the way for a more data-driven approach to healthcare.

Author contributions

All authors wrote the various sections of the initial drafted manuscript and S. Albahra and H. Rashidi critically reviewed and edited the drafts. All authors read and commented on the paper and approved submission.

Funding

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

Declaration of Competing Interest

MILO (the auto-ML platform mentioned) is the intellectual property of the Regents of the University of California (UC) and two of the co-authors in this manuscript (H. Rashidi & S. Albahra) are its co-inventors. Both are also on the board of MILO-ML Inc. (a UC start up). The other authors have no conflict of interests to declare

Acknowledgements

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

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1053/j.semdp.2023.02.002.

References

1 Rashidi HH, Albahra S, Robertson S, Tran NK, Hu B. Common statistical concepts in the supervised machine learning arena. *Front Oncol*. 2023;13. Accessed February 10, 2023 https://www.frontiersin.org/articles/10.3389/fonc.2023.1130229.
2 NLP - overview. Accessed October 10, 2022. https://cs.stanford.edu/people/eroberts/courses/soco/projects/2004-05/nlp/overview_history.html.
3 Chomsky N. Three models for the description of language. *IEEE Trans Inf Theory*. 1956;2(3):113–124. https://doi.org/10.1109/TIT.1956.1056813.
4 Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544–551. https://doi.org/10.1136/amiajnl-2011-000464.
5 Falconer W. An Universal Dictionary of the Marine: Or, A Copious Explanation of the Technical Terms and Phrases Employed in the Construction, Equipment, Furniture, Machinery, Movements, and Military Operations of a Ship. *T. Cadell*; 1784. https://books.google.com/books?id=3pVAAAAAYAAJ.
6 Chapman W, Dowling J, Chu D. *ConText: an algorithm for identifying contextual features from clinical text. Biological, Translational, and Clinical Language Processing*. Association for Computational Linguistics; 2007:81–88. Accessed October 19, 2022 https://aclanthology.org/W07-1011.
7 Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than English: opportunities and challenges. *J Biomed Semant*. 2018;9(1):12. https://doi.org/10.1186/s13326-018-0179-8.
8 Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural language processing in oncology: a review. *JAMA Oncol*. 2016;2(6):797–804. https://doi.org/10.1001/jamaoncol.2016.0213.
9 Ranganathan P, Gogtay NJ. An introduction to statistics - data types, distributions and summarizing data. *Indian J Crit Care Med Peer-Rev Off Publ Indian Soc Crit Care Med*. 2019;23(Suppl 2):S169–S170. https://doi.org/10.5005/jp-journals-10071-23198.
10 Bensken WP, Pieracci FM, Ho VP. Basic introduction to statistics in medicine, part 1: describing data. *Surg Infect*. 2021;22(6):590–596. https://doi.org/10.1089/sur.2020.429.
11 Glen S. Poisson distribution /Poisson curve: simple definition. Statistics How To. Published 2018. Accessed October 10, 2022. https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/probability-distribution/poisson-distribution/.
12 Starkhagen C. Qualitative data: the unsung hero of machine learning datasets. Published January 18, 2022. https://www.twine.net/blog/qualitative-data-machine-learning-datasets/.
13 Wang Y, Ge X, Ma H, Qi S, Zhang G, Yao Y. Deep learning in medical ultrasound image analysis: a review. *IEEE Access*. 2021;9:54310–54324. https://doi.org/10.1109/ACCESS.2021.3071301.
14 Wright AI, Dunn CM, Hale M, Hutchins GGA, Treanor DE. The effect of quality control on accuracy of digital pathology image analysis. *IEEE J Biomed Health Inform*. 2021;25(2):307–314. https://doi.org/10.1109/JBHI.2020.3046094.
15 Juluru K, Shih HH, Keshava Murthy KN, Elnajjar P. Bag-of-words technique in natural language processing: a primer for radiologists. *Radiogr Rev Publ Radiol Soc N Am Inc*. 2021;41(5):1420–1426. https://doi.org/10.1148/rg.2021210025.
16 Lokesh S, Devi MR. Speech recognition system using enhanced mel frequency cepstral coefficient with windowing and framing method. *Clust Comput*. 2019;22(5):11669–11679. https://doi.org/10.1007/s10586-017-1447-6.
17 Multicollinearity assessment & removal tool | machine intelligence learning optimizer (MILO-ML) documentation (v2.2.1). Accessed October 14, 2022. https://milo-ml.com/docs/processor-guide/multicollinearity.html.
18 Păpăluță V. What's the best way to handle NaN values? Medium. Published January 14, 2020. Accessed October 13, 2022. https://towardsdatascience.com/whats-the-best-way-to-handle-nan-values-62d50f738fc.
19 Imputation & encoder tool (MILO Pro) | Machine intelligence learning optimizer (MILO-ML) documentation (v2.2.1). Accessed November 2, 2022. https://milo-ml.com/docs/processor-guide/imputation-encoder.html.
20 Brownlee J. Why one-hot encode data in machine learning? Machine Learning Mastery. Published June 30, 2020. Accessed October 13, 2022. https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/.
21 Al-Zebari A, Sengur A. Performance comparison of machine learning techniques on diabetes disease detection. In: *2019 1st International Informatics and Software Engineering Conference (UBMYK)*. 2019:1–4. https://doi.org/10.1109/UBMYK48245.2019.8965542.
22 Jayatilake SMDAC, Ganegoda GU. Involvement of machine learning tools in healthcare decision making. *J Healthc Eng*. 2021;2021, e6679512. https://doi.org/10.1155/2021/6679512.
23 Michalski RS, Carbonell JG, Mitchell TM. *Machine Learning: An Artificial Intelligence Approach*. Springer Science & Business Media; 2013.
24 Cárdenas-López FA, Lamata L, Retamal JC, Solano E. Multiqubit and multilevel quantum reinforcement learning with quantum technologies. *PLOS ONE*. 2018;13(7), e0200455. https://doi.org/10.1371/journal.pone.0200455.
25 Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. *Acad Pathol*. 2019;6, 2374289519873088. https://doi.org/10.1177/2374289519873088.
26 Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLOS Med*. 2018;15(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689.
27 Arbet J, Brokamp C, Meinzen-Derr J, Trinkley KE, Spratt HM. Lessons and tips for designing a machine learning study using EHR data. *J Clin Transl Sci*. 2021;5(1):e21. https://doi.org/10.1017/cts.2020.513.
28 Rashidi HH, Bowers KA, Gil MR. Machine learning in the coagulation and hemostasis arena: an overview and evaluation of methods, review of literature, and future directions. *J Thromb Haemost*. 2022;0(0). https://doi.org/10.1016/j.jtha.2022.12.019.
29 Yang Q., Liu Y., Chen T., Tong Y. Federated machine learning: Concept and Applications. Published online February 13, 2019. doi:10.48550/arXiv.1902.04885.
30 Duckworth C, Chmiel FP, Burns DK, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. *Sci Rep*. 2021;11(1):23017. https://doi.org/10.1038/s41598-021-02481-y.
31 Harrison JH, Gilbertson JR, Hanna MG, et al. Introduction to artificial intelligence and machine learning for pathology. *Arch Pathol Lab Med*. 2021;145(10):1228–1254. https://doi.org/10.5858/arpa.2020-0541-CP.
32 LeCun Y, Haffner P, Bottou L, Bengio Y. Object recognition with gradient-based learning. In: Forsyth DA, Mundy JL, di Gesú V, Cipolla R, eds. *Shape, Contour and Grouping in Computer Vision. Lecture Notes in Computer Science*. Springer; 1999:319–345. https://doi.org/10.1007/3-540-46805-6_19.
33 Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks. *Pattern Recognit*. 2018;77:354–377. https://doi.org/10.1016/j.patcog.2017.10.013.
34 Boureau Y.L., Ponce J., LeCun Y. A theoretical analysis of feature pooling in visual recognition.:8.
35 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014:580–587. https://doi.org/10.1109/CVPR.2014.81.

36 Ciresan D, Giusti A, Gambardella L, Schmidhuber J. *Deep neural networks segment neuronal membranes in electron microscopy images*. In: *Advances in Neural Information Processing Systems*. 25. Curran Associates, Inc.; 2012. Accessed October 27, 2022 https://proceedings.neurips.cc/paper/2012/hash/459a4ddcb586f24efd9395aa7662bc7c-Abstract.html.

37 Asif H, Vaidya J, Shafiq B, Adam N. Secure and efficient k-NN queries. In: De Capitani di Vimercati S, Martinelli F, eds. *ICT Systems Security and Privacy Protection. IFIP Advances in Information and Communication Technology*. Springer International Publishing; 2017:155–170. https://doi.org/10.1007/978-3-319-58469-0_11.

38 Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat*. 1992;46(3):175–185. https://doi.org/10.1080/00031305.1992.10475879.

39 Hall P, Park BU, Samworth RJ. Choice of neighbor order in nearest-neighbor classification. *Ann Stat*. 2008;36(5):2135–2152. https://doi.org/10.1214/07-AOS537.

40 SEAL HL. Studies in the history of probability and statistics. XV The historical development of the Gauss linear model. *Biometrika*. 1967;54(1–2):1–24. https://doi.org/10.1093/biomet/54.1-2.1.

41 Schober P, Vetter TR. Linear regression in medical research. *Anesth Analg*. 2021;132(1):108–109. https://doi.org/10.1213/ANE.0000000000005206.

42 Aggarwal R, Ranganathan P. Common pitfalls in statistical analysis: the use of correlation techniques. *Perspect Clin Res*. 2016;7(4):187–190. https://doi.org/10.4103/2229-3485.192046.

43 Predictive modelling using linear regression | by RAJAT PANCHOTIA | The startup | Medium. Accessed November 2, 2022. https://medium.com/swlh/predictive-modelling-using-linear-regression-e0e399dc4745.

44 Bisong E. Logistic regression. In: Bisong E, ed. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. Apress; 2019:243–250. https://doi.org/10.1007/978-1-4842-4470-8_20.

45 Support Vector Machines Part 1 (of 3): Main Ideas!!!; 2019. Accessed October 18, 2022. https://www.youtube.com/watch?v=efR1C6CvhmE.

46 Uddin S, Ong S, Lu H. Machine learning in project analytics: a data-driven framework and case study. *Sci Rep*. 2022;12(1):15252. https://doi.org/10.1038/s41598-022-19728-x.

47 Hyafil L, Rivest RL. Constructing optimal binary decision trees is NP-complete. *Inf Process Lett*. 1976;5(1):15–17. https://doi.org/10.1016/0020-0190(76)90095-8.

48 Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1(1):81–106. https://doi.org/10.1007/BF00116251.

49 Papagelis A, Kalles D. Breeding decision trees using evolutionary techniques. In: *Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01*. Morgan Kaufmann Publishers Inc.; 2001:393–400.

50 Mehta D, Raghavan V. Decision tree approximations of Boolean functions. *Theor Comput Sci*. 2002;270(1):609–623. https://doi.org/10.1016/S0304-3975(01)00011-1.

51 Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189–1232.

52 Rahman S, Irfan M, Raza M, Moyeezullah Ghori K, Yaqoob S, Awais M. Performance analysis of boosting classifiers in recognizing activities of daily living. *Int J Environ Res Public Health*. 2020;17(3):1082. https://doi.org/10.3390/ijerph17031082.

Further reading

1 Chen PH. Essential elements of natural language processing: what the radiologist should know. *Acad Radiol*. 2020;27(1):6–12. https://doi.org/10.1016/j.acra.2019.08.010.

2 Brownlee J. *A gentle introduction to the bag-of-words model. Machine Learning Mastery*; 2017. Published October 8,Accessed October 19, 2022 https://machinelearningmastery.com/gentle-introduction-bag-words-model/.