AJR FOCUS ON:

# Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods

Guy S. Handelman[1,2]
Hong Kuan Kok[3,4]
Ronil V. Chandra[5,6]
Amir H. Razavi[7,8]
Shiwei Huang[9]
Mark Brooks[5,10]
Michael J. Lee[2,11]
Hamed Asadi[5,10,12]

**Keywords:** artificial intelligence, machine learning, medicine, supervised machine learning, unsupervised machine learning

doi.org/10.2214/AJR.18.20224

Received June 6, 2018; accepted after revision August 2, 2018.

[1]Department of Radiology, Belfast City Hospital, 51 Lisburn Rd, Belfast, Antrim BT9 7AB, UK. Address correspondence to G. S. Handelman (guyhandelman@rcsi.ie).

[2]Royal College of Surgeons in Ireland, Dublin, Ireland.

[3]Interventional Radiology Service, Northern Hospital Radiology, Epping, Australia.

[4]School of Medicine, Faculty of Health, Deakin University, Waurn Ponds, Australia.

[5]Interventional Neuroradiology Service, Monash Imaging, Monash Health, Clayton, Australia.

[6]Faculty of Medicine, Nursing and Health Sciences, Monash University, Clayton, Australia.

[7]School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada.

[8]BCE Corporate Security, Ottawa, ON, Canada.

[9]The Australian National University Medical School, Garran, Australia.

[10]Department of Radiology, Interventional Neuroradiology Service, Austin Health, Heidelberg, Australia.

[11]Department of Radiology, Beaumont Hospital, Dublin, Ireland.

[12]The Florey Institute of Neuroscience and Mental Health, University of Melbourne, Australia.

This article is available for credit.

*AJR* 2019; 212:38–43

0361–803X/19/2121–38



**OBJECTIVE.** Machine learning (ML) and artificial intelligence (AI) are rapidly becoming the most talked about and controversial topics in radiology and medicine. Over the past few years, the numbers of ML- or AI-focused studies in the literature have increased almost exponentially, and ML has become a hot topic at academic and industry conferences. However, despite the increased awareness of ML as a tool, many medical professionals have a poor understanding of how ML works and how to critically appraise studies and tools that are presented to us. Thus, we present a brief overview of ML, explain the metrics used in ML and how to interpret them, and explain some of the technical jargon associated with the field so that readers with a medical background and basic knowledge of statistics can feel more comfortable when examining ML applications.

**CONCLUSION.** Attention to sample size, overfitting, underfitting, cross validation, as well as a broad knowledge of the metrics of machine learning, can help those with little or no technical knowledge begin to assess machine learning studies. However, transparency in methods and sharing of algorithms is vital to allow clinicians to assess these tools themselves.

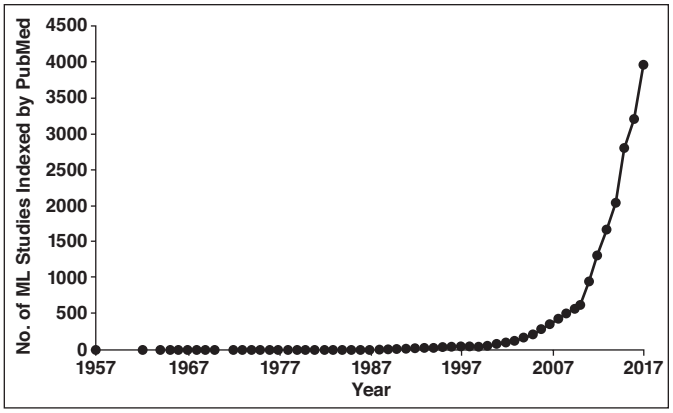
Machine learning (ML) is the application of computing power and algorithms coupled with the ability for the algorithm to learn from data and experience. This description is somewhat broad, akin to saying statistics is the application of mathematics to datasets. The reality is more complex, and ML applications are heterogeneous and varied. The recent prominence of ML in medicine is because of a variety of factors such as the availability of cheaper and more powerful computing, larger datasets, and applications in nonmedical fields. These factors have led to an almost exponential rise in the number of ML studies published in the literature in recent years (Fig. 1) and to an increased focus on ML at conferences. At the Radiological Society of North America 2017 annual meeting, there were 49 ML or artificial intelligence (AI) exhibitors showcasing practical applications of ML to health care. These exhibitors were not academic showcases showing theory but were private companies already applying ML in meaningful ways, which shows that ML has transitioned from the academic sphere to day-to-day reality in health care enough to attract venture capital. The big names in the information technology industry have not missed this trend; Microsoft, Google, and Amazon are making forays into the area by both contributing to ML tools and conducting primary research into the application of these tools [1].

The roots of ML can be traced back to the 1950s when Alan Turing developed the so-called "learning machine" and military applications of basic AI [2]. At that time, computers were huge and storage capacity was prohibitively expensive; thus, the capabilities—although significant for that time—were limited. Over the following decades, stepwise improvements in theory and technologic advances steadily increased the power and versatility of ML, and its ability to deal with data brought it to the attention of the classical statistics community. Conversely, the computer science discipline saw how tried and tested algorithms and statistical processes of classical statistics could be incorporated into ML [3]. However, this convergence of cultures has led to two sets of nomenclature being merged into the field of ML, and thus confusion can arise from this new terminology. Most readers will be familiar with the concept of a *t* test or a *p* value, but some will not understand the term "area under the ROC curve" or "AUROC" [4]. Although there are a number of articles in the literature that give overviews of

ML as it applies to medicine and radiology, many are overly complicated for the novice reader and focus on the intricacies of the various models and their applications as opposed to educating readers on how to evaluate what the metrics and results mean. For a more advanced introduction to the field and expansion on many of the points raised here, we recommend the "Points of Significance" series in *Nature Methods*, which provides an accessible and cogent exploration of many core elements of ML [5].

Within medicine, most ML applications are focused on prognostication (predicting outcomes), categorization (assigning patients to groups), detection (identifying outliers or abnormal findings), and dimensionality reduction (reducing the number of uninformative variables or elucidating the variables that are the most informative) [6]. Although the number of ML algorithms is vast and description of each one is beyond the scope of this article, most ML algorithms fall into two broad categories: supervised learning and unsupervised learning. These terms refer to whether the user supervises the program when it is making its predictive algorithm. In supervised learning, the user supervises by providing the program with solved problems; for example, a user provides a database of 1000 patients in whom it is known whether they develop a condition or not (e.g., large-vessel occlusion in the cerebral circulation). The program can then use a variety of algorithms to determine which patient variables (e.g., body mass index [BMI], blood pressure, biochemical markers, CT images) in what combination predicts the progression to a disease state [7]. The program learns which variables predict disease and can then predict or diagnose a condition in new unseen patients (Fig. 2). Subsequently, once the patients develop a condition (or not), the program can then use these data to improve its predictive capability and iteratively improve its performance.

**Fig. 1**—Number of machine learning (ML) studies indexed by PubMed (U.S. National Library of Medicine) per year.




With unsupervised learning, by comparison, the ML algorithm is not explicitly instructed about what type of answer to produce; instead, it is tasked with determining whether latent patterns within data exist [8, 9]. For our previous example of patients with known presence or absence of a large-vessel occlusion in the cerebral circulation, unsupervised learning is concerned with determining if there are any particular patterns in the patient data that would point toward certain subpopulations who are higher risk or perhaps if there is a trend in the myriad patient variables that seems to contribute more than others toward determining disease occurrence, such as the combination of low blood pressure, tachycardia, and a raised sodium level in the blood.

These processes of supervised and unsupervised learning can be blended to another process called "semisupervised learning." As the name implies, semisupervised learning is when there is a mix of solved or labeled data and unlabeled data. Semisupervised learning is of particular utility in medicine and radiology because there frequently are vast amounts of unlabeled data that would be overly onerous to manually label. As an example, if attempting to train an algorithm to detect, segment, and diagnose subtypes of malignancies from cross-sectional imaging, the label would extend usually only to the final diagnosis; slice-by-slice and voxel-specific annotations would be missing, and adding these annotations would be extremely taxing. However, if only a few highly labeled images were added to the dataset, overall performance of the ML algorithm can be increased [10].

There are a multitude of ways that ML has been applied to medical questions with impressive results; however, detailing these applications is beyond the scope of this article. Instead, we will focus on how to assess the metrics of ML so that readers will be more familiar with the terms used in these types of research and what they mean.

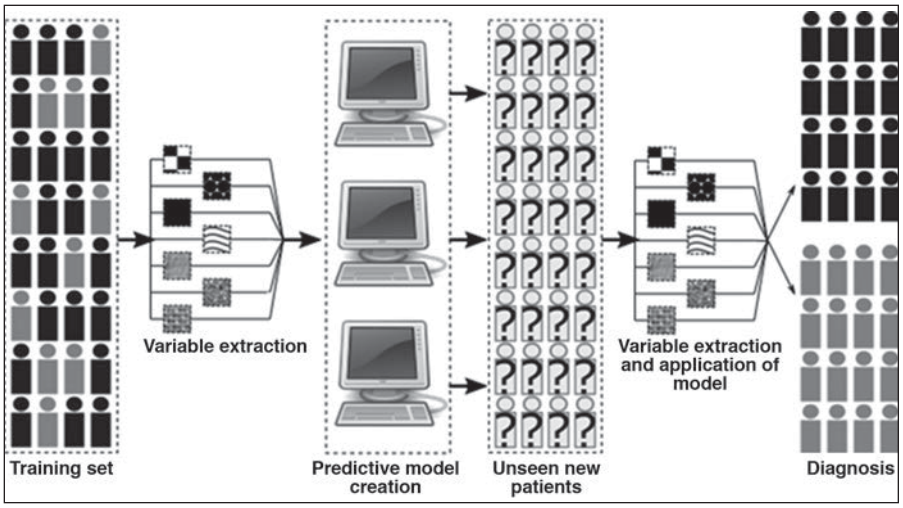



**Fig. 2**—Schematic shows basic concept of supervised machine learning (ML) applied to predicting disease occurrence. Training set of patients with presence (*black*) or absence (*gray*) of disease is predicted by learning which combination of variables can produce model that can predict disease presence when applied to new patients in whom diagnosis is unknown (*question marks*).

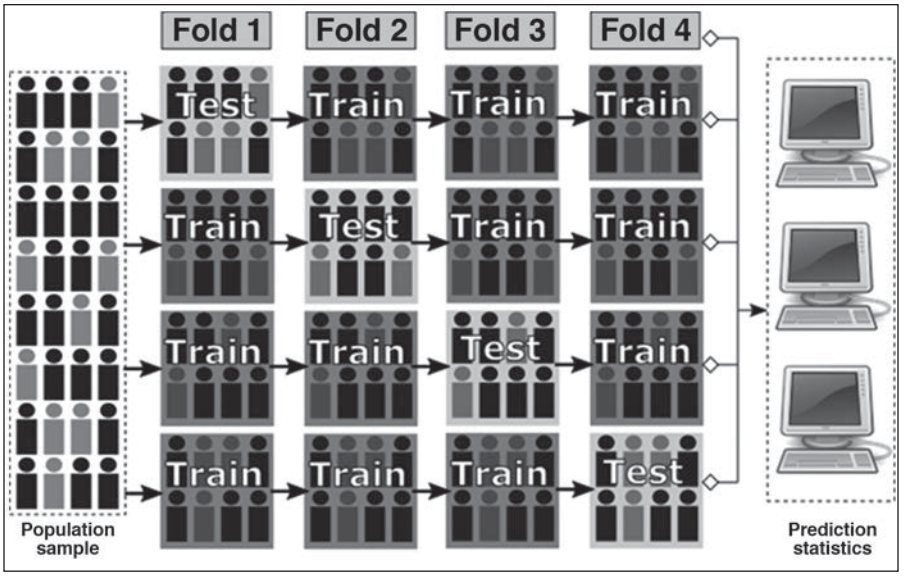



**Fig. 3**—Schematic shows simplified cross-validation technique. Population sample is composed of patients with (*black*) and without (*gray*) disease. Dataset is partitioned into series of training and test sets (folds), and cumulative mean of metrics allows more robust assessment of model performance.

## Metrics and Assessment of Machine Learning Algorithms

Claims of the effectiveness of ML algorithms often detail the quality of the algorithm using a core set of performance measurements. Most medical readers will be comfortable with the *p* value statistic and understand what it means when results are reported to be greater or less than significance or less than 0.05. However, some readers will not be familiar with the term "area under the ROC curve" or "AUROC." If current trends continue, understanding terms such as the latter one will become increasingly important for all practitioners in every discipline of medicine.

If dealing with ML models that classify (e.g., predict the presence or absence of a medical condition, predict which patients will have good vs bad outcome), then models are usually evaluated with the ROC curve or confusion matrix [11, 12]. If a model deals with regression (e.g., predicting response or outcomes using continuous variables such as predicting longevity in months for patients with lung cancer on the basis of tumor burden on CT combined with patient comorbidities), then the model is usually evaluated with mean squared error (MSE), mean absolute error (MAE), or the coefficient of determination ($R^2$). To an extent, these terms are technical jargon for easy-to-understand concepts, and readers with advanced statistical knowledge will recognize significant corollaries to a classical statistical approach. Before we describe these terms in detail, we must first discuss the concept of cross validation, which is the method by which many ML algorithms begin to generate performance measures.

## Cross Validation

To test whether a predictive algorithm works, a researcher must test it on a population. Instead of creating a predictive algorithm and testing it on a new population, many researchers will take an initial group of subjects and partition them into a training dataset and a testing dataset. The training dataset is composed of subjects that are used to create the algorithm that will perform the predictions. Using the example mentioned earlier, if one had a database of 1000 patients with suspected ischemic stroke who were subsequently proved to have or not have a proximal intracranial vascular occlusion, then one could use 800 patients to train the algorithm and the remaining 200 patients as unseen patients to test the algorithm as if they were new patients presenting to the hospital. This can be taken one step further and can avoid the cost of


**Fig. 4**—Overfitting and underfitting.
**A–C,** When separating data into categories (**A**), it is possible to overfit (**B**) model to sample data, making generalization of model to population unrealistic. Similarly, by making model very generalizable (**C**), it is possible to underfit model and make predictions weak. Dotted line = calculated decision or prediction line to separate data into groups, gray = condition present, black = condition absent.

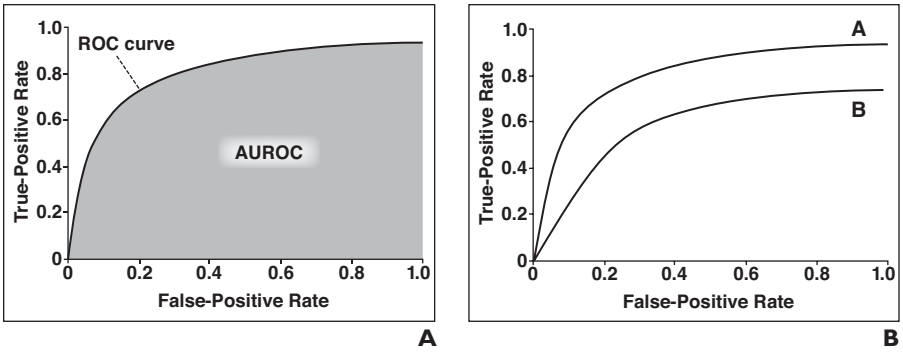



**Fig. 5**—Area under ROC curve (AUROC).
**A,** AUROC is function of effect of different sensitivities (true-positive rate) on rate of false-positives.
**B,** By comparing AUROC values, we can quickly see that ROC curve A has greater AUROC than ROC curve B and thus indicates better overall performance.

limiting the size of the dataset by repeating this process many times, each time assigning a different group of study patients to the training set and to the testing set [13] (Fig. 3). Each iteration will not only improve the performance of the model, because the program can compare between each training set's results to see what performs best and can alter its overall predictive capability, but also improve the generalizability of the results. As an algorithm deals with many combinations of patients, the chance of overfitting the predictive algorithm decreases (i.e., being exceptionally good at predicting results on the basis of a training set but less reliable in a general population because the algorithm has learned the variations of the training set too well) [13] (Fig. 4).

This process of validating an ML algorithm via demonstration on a theoretic test set allows the numeric representation of an algorithm's performance because it is able to see how many predictions were correct (true-positives and true-negatives) and incorrect (false-positives and false-negatives); how precise (how much variability repeated predictions show compared with the true values [if all are grouped closely together, they are more precise]) or accurate (how close to the true value predictions lie, also called bias) the predictions are; and, by extension, the sensitivity and specificity. In ML, these metrics are commonly summarized in a number of ways for a variety of reasons. With modern algorithms, one can choose how much weight is applied toward sensitivity or specificity; ideally, both will be high, but frequently, to raise sensitivity, a compromise in specificity must be made. This concept has led to the much-used ROC curve [14].

## ROC Curve

By plotting the effect of different levels of sensitivity on specificity, a curve can be made that represents the performance of a particular predictive algorithm, thereby allowing readers to quickly understand the utility of the algorithm. For different tasks, a particular operating point on the curve can be used. For example, if screening for large areas of ischemia in a brain on CT that would then be reviewed by a radiologist, one might want a higher sensitivity, accepting that there will be a percentage of false-positives preoccupying the interpreter. However, if there are too many false-positives, then the sensitivity can be augmented so that a more acceptable burden of false-positives is presented. Deciding which levels to use are task- and system-specific, with many applications in tasks such as CT detection of colonic polyps or lung nodules [15–17]. This plot of the true-positive rate versus the false-positive rate is called the "ROC curve" (Fig. 5), and a function of the ROC curve is calculating the area beneath the ROC curve or AUROC. Algorithms that perform better will have a higher sensitivity and specificity and thus the area under the plotted line will be greater than those that perform worse. The metric termed the "area under the ROC curve" or "AUROC" is commonly quoted and offers a quick way to compare algorithms. Many investigators of published studies will select which point on the ROC curve that they believe will provide the best results for the task.

For tasks that require identification and localization of abnormalities on an image, ROC curves evaluate detection accuracy but do not evaluate whether the algorithm correctly identifies the locations of abnormalities. Thus, for these tasks, the free ROC (FROC) curve is used. FROC curves are similar to ROC curves but use localization accuracy and, in cases in which there are many abnormal areas, the fractions of true-positives and false-positives are calculated. Interpretation of the acceptable distance from the true lesion when defining localization accuracy is usually set by the investigator before an FROC analysis; thus, if the task requires tight localization (e.g., in acute isch-

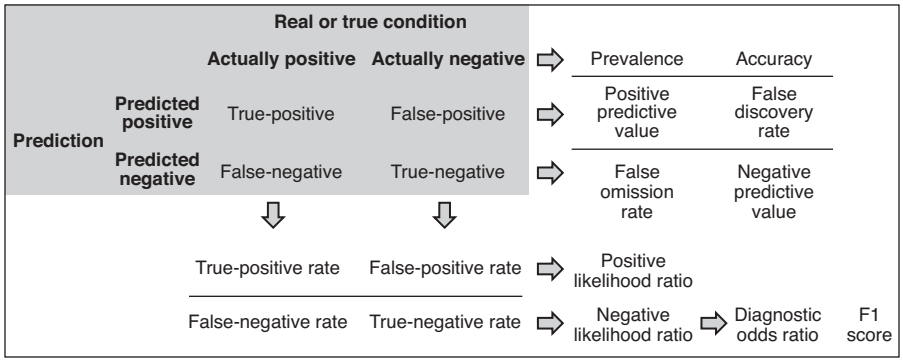



**Fig. 6**—Components of confusion matrix. True-positive rate is also known as recall or sensitivity and is probability of model detecting truly positive case. False-negative rate is proportion of positive cases missed. False-positive rate or fallout is probability of incorrectly identifying case as being positive when condition is not actually present. True-negative rate or specificity is probability of being correct in ruling out condition. Accuracy is amount of correct calls (true-positive and true-negative) that were made in proportion to total dataset. Positive predictive value or precision is likelihood that positive prediction is actually positive; similarly, negative predictive value is likelihood negative prediction is actually negative. Positive and negative likelihood ratios are probabilities that if test results for individual are positive or negative that individual is truly positive or negative for disease. Diagnostic odds ratio is measurement of effectiveness of test but, unlike accuracy, is independent of prevalence. F1 score is function of true-positive rate and positive predictive value (precision and recall) to give overall indication of performance of classifier.

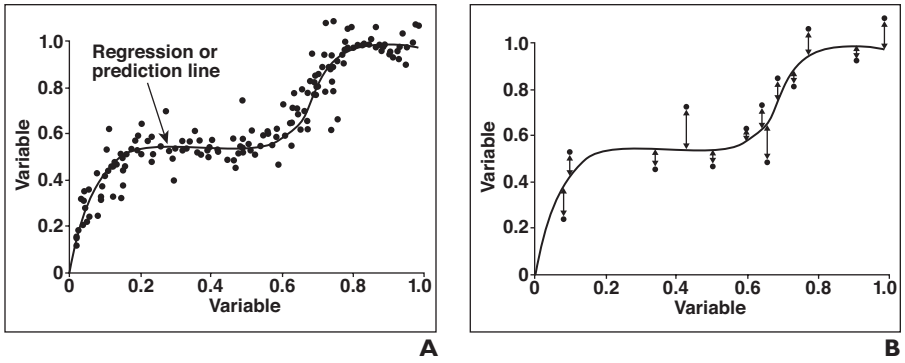



**Fig. 7**—Calculating error of regression model.
**A,** Training dataset. Predictor that predicts change in desired outcome on basis of input variables in training dataset can be created. Black dots represent data points on theoretical *xy*-axis.
**B,** Test dataset. Once predictor is applied to test dataset, measurement of error (*arrows*) is possible.



emic stroke) and the acceptable distance is set inappropriately high, the algorithm could be made to seem to be performing better than it would in real clinical tasks [18]. Although FROC curves resemble ROC curves, the added levels of calculation and dependency on the sample-specific disease locations mean that a direct numeric representation of performance such as the AUROC is difficult, and the use of fractions means that the FROC curve may have a large horizontal axis.

## Confusion Matrix

In discussing AUROC, we touched briefly on a few statistical terms that are used, such as sensitivity and false-positive rate. Many different terms are used in varying circumstances in the literature. However, an understanding of the confusion matrix clarifies many of these terms. In its basic form, the confusion matrix is a representation of many of these terms, so readers who are interested in a specific metric of a particular algorithm can quickly see that metric and compare that algorithm with other algorithms. By comparing true-positives and true-negatives in a test sample with the predicted positive and predicted negative data, a variety of summary statistics can be calculated [19] (Fig. 6). Many studies will include a limited form of this information—for example, showing only the true- and false-positives and true- and false-negatives, the predicted positives and negatives, and actual positives and negatives; the remaining metrics can be calculated from these values if readers want to compare different metrics between models.

One value that is usually quoted prominently in published studies is accuracy, which is the number of correct predictions made as a ratio to all predictions made; higher is better. A model that detects the presence or absence of a significant pulmonary embolus on pulmonary angiography with an accuracy of 90% is wrong in 10% of the cases; however, incorrectly identifying a pulmonary embolus as being absent is probably worse than saying it is present. Thus, depending on the clinical question, it may be more important to consider the other metrics of the confusion matrix such as the negative likelihood value, which informs the clinician more directly than accuracy about how reliable a negative test result or prediction is.

## Mean Squared Error and Mean Absolute Error

In applications of ML to problems of regression, the relationship between variables is elucidated by creating an equation that minimizes the distance between a fitted line and the data point and, by extension, predictions can be made. A measure of the degree that the regression line fits the data and thus reliably makes predictions is represented by the MSE. The MSE is usually calculated by applying the equation for the line of regression to the set of known variables to see how much variance from the regression line results. For the purposes of model evaluation, one can apply the regression line formed from learning from a training dataset to make predictions in a test dataset and see how much the predicted variables differ from the actual variables [20] (Fig. 7). Thus, the error is the mean deviation from the true values; lower is better. If you created a predictor that calculated the modified Rankin score after intervention for acute ischemic stroke on the basis of premorbid variables, such as BMI, blood pressure, and comorbidities (e.g., diabetes), and applied it to a test dataset, then the amount of under- or overestimation of the modified Rankin score based on these variables is the error. Multiple variants of this error calculation exist: the MSE, the root mean squared error (RMSE), the MAE, and the mean absolute percentage error (MAPE). All differ slightly in what they represent and, strictly speaking, are not directly comparable. For example, MSE is better metric if having large errors is undesirable. However, all share a common component: Smaller is better. An exception is the commonly used coefficient of determination ($R^2$) metric, which is a similar measure of goodness of fit of the model to the data (i.e., how much of the observed variations in the data are explained by the model); in the case of $R^2$ values, 0% is worst and 100% is best [21].

## Image Segmentation Evaluation

For detection tasks for images, it is sometimes useful for an ML algorithm to be accurate in determining the specific localization of anatomy or pathologic findings—for example, in determining the size and extent of a tumor or for image segmentation. Assessing ML algorithms for their accuracy in these tasks does not fall into a straightforward true-positive versus false-positive evaluation because the accuracy of the image registration is of interest as well. Thus, using metrics such as ROC, FROC, and MSE is inappropriate. Many investigators use metrics to show how good or bad an ML algorithm's predictions are, and there are a variety of these metrics (e.g., intersection over union, Dice coefficient, Jaccard index); however, all are more or less based on the same principle and have the same numeric representation. In this evaluation method, the predicted area of interest generated by the algorithm is compared against an ideal or completely accurate evaluation image. Taking our previous example as an illustration: If an ML algorithm detects a brain tumor on a set of cross-sectional images of the brain and graphically outlines the tumor extent, this can be compared with a solved case in which a human operator outlined the tumor on identical images. The degree of overlap of the two is taken as a representation of accuracy of the model with values ranging from 0 to 1 (perfect agreement). Depending on the task and accuracy needed, different values are acceptable;

however, in general, a score should be greater than 0.5 to be considered. This can be taken one step further in combining image segmentation accuracy with precision and recall rates because it is important that the algorithm not only correctly identifies the boundaries of a pathologic finding, but also correctly identifies a pathologic finding as being present or absent. This is commonly done by representing these combined metrics as the mean average precision (MAP). The MAP is calculated by averaging the precision and recall rates at different thresholds of segmentation accuracy. By increasing the threshold of segmentation accuracy (i.e., forcing the algorithm to accept only a high degree of overlap before it can define a pathologic finding as present), there will usually be a trade-off in precision and recall. Similarly, if the algorithm is designed to detect more than one type of pathologic finding or anatomy (e.g., detect, classify, and outline different subtypes of brain tumors on cross-sectional imaging), it may perform better on one type of pathologic finding than another; thus, MAP can be used to average its performance across all tasks.

## Conclusion

For those who are unfamiliar with the field of ML, the emerging research can be daunting, with a wide variation in the terms used and the metrics presented. How can we, as readers, tell if the predictive model being presented is good or is even better than another model presented elsewhere? In broad terms, in classification problems with metrics such as AUROC and accuracy, it is clear that higher is better. In regression problems, lower error is better, and a higher $R^2$ score is better. However, readers must be careful to discern what is being compared. A small sample size—irrespective of the mathematic manipulations applied—is not comparable to a larger sample size. If you have a small training dataset and an even smaller testing set, you can create a "perfect" prediction model for your dataset. However, the utility of ML models is their generalizability. If a model cannot perform in real-world scenarios, then it is purely an academic exercise. Therefore, it is important that databases are expanded and maintained with open access to data and the predictive algorithm so that reported performance can be verified and competing algorithms are drawing from the same data pool. If one predictive tool is trained on a separate dataset from another predictive tool, true comparisons are hard to achieve. Because the end goal of creating assistive technology—that is, to improve service and patient outcomes—depends on clinicians actually using these tools, we advocate the standard that research studies of ML algorithms for medical applications should include the end product, the developed predictive algorithm.

## References


1. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; 316:2402–2410
2. Turing AM. Computing machinery and intelligence. *Mind* 1950; 59:433–460
3. Patel VL, Shortliffe EH, Stefanelli M, et al. The coming of age of artificial intelligence in medicine. *Artif Intell Med* 2009; 46:5–17
4. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci* 2001; 16:199–231
5. Lever J, Krzywinski M, Altman N. Classification evaluation. *Nat Methods* 2016; 13:603–604
6. Deo RC. Machine learning in medicine. *Circulation* 2015; 132:1920–1930
7. Asadi H, Dowling R, Yan B, Mitchell P. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS One* 2014; 9:e88225
8. Guan WJ, Jiang M, Gao YH, et al. Unsupervised learning technique identifies bronchiectasis phenotypes with distinct clinical characteristics. *Int J Tuberc Lung Disease* 2016; 20:402–410
9. Sajda P. Machine learning for detection and diagnosis of disease. *Annu Rev Biomed Eng* 2006; 8:537–565
10. Peikari M, Salama S, Nofech-Mozes S, Martel AL. A cluster-then-label semi-supervised learning approach for pathology image classification. *Sci Rep* 2018; 8:7193
11. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manage* 2009; 45:427–437
12. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform* 2005; 38:404–415
13. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Stat Surv* 2010; 4:40–79
14. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143:29–36
15. Summers RM, Yao J, Pickhardt PJ, et al. Computed tomographic virtual colonoscopy computer-aided polyp detection in a screening population. *Gastroenterology* 2005; 129:1832–1844
16. Wang S, Summers RM. Machine learning and radiology. *Med Image Anal* 2012; 16:933–951
17. Chan HP, Hadjiiski L, Zhou C, Sahiner B. Computer-aided diagnosis of lung cancer and pulmonary embolism in computed tomography: a review. *Acad Radiol* 2008; 15:535–555
18. Moskowitz CS. Using free-response receiver operating characteristic curves to assess the accuracy of machine diagnosis of cancer. *JAMA* 2017; 318:2250–2251
19. Ting KM. Confusion matrix. In: Sammut C, Webb GI, eds. *Encyclopedia of machine learning*. Boston, MA: Springer, 2010:209
20. Wallach D, Goffinet B. Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecol Modell* 1989; 44:299–306
21. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning: with applications in R*. New York, NY: Springer Publishing, 2014:430