Critical Appraisal of a Machine Learning Paper: A Guide for the Neurologist

Pulikottil W. Vinny, Rahul Garg[1], MV Padma Srivastava[2], Vivek Lal[3], Venugoapalan Y. Vishnu[4]

Neurology, Indian Naval Hospital Ship Asvini, Mumbai, [1]Department of Computer Science and Engineering, Indian Institute of Technology, Delhi, [2]Neurology, and Chief of Neurosciences Centre, All India Institute of Medical Sciences, New Delhi, [3]Neurology, Postgraduate Institute of Medical Education and Research, Chandigarh, [4]Neurology, AIIMS New Delhi, India

Abstract

Machine learning (ML), a form of artificial intelligence (AI), is being increasingly employed in neurology. Reported performance metrics often match or exceed the efficiency of average clinicians. The neurologist is easily baffled by the underlying concepts and terminologies associated with ML studies. The superlative performance metrics of ML algorithms often hide the opaque nature of its inner workings. Questions regarding ML model's interpretability and reproducibility of its results in real-world scenarios, need emphasis. Given an abundance of time and information, the expert clinician should be able to deliver comparable predictions to ML models, a useful benchmark while evaluating its performance. Predictive performance metrics of ML models should not be confused with causal inference between its input and output. ML and clinical gestalt should compete in a randomized controlled trial before they can complement each other for screening, triaging, providing second opinions and modifying treatment.

Keywords: Critical appraisal, deep learning, machine learning, neural networks

Introduction

Machine learning (ML), a form of artificial intelligence (AI), is being increasingly employed in neurology.[1-4] A key feature behind the excitement in ML is its potential to analyze large and complex data structures (big data) to create prediction models that promise to improve diagnosis, prognosis, monitoring, and administration of treatments. ML has been employed in diagnosing stroke from neuroimaging, detecting papilledema, and diabetic retinopathy (DR) from retinal scans, interpreting electroencephalogram to prognosticate coma, detecting seizure before ictus, predicting conversion of mild cognitive impairment to Alzheimer's dementia, and classifying neurodegenerative diseases based on gait and handwriting.[5-14] The reported performance metrics of these ML studies match or exceed the efficiency of the average clinicians. The eager neurologist often struggles to get a grip on the concepts behind the stellar performance of ML tools. In this review, we try to simplify the process of critical appraisal when reading a research paper which uses ML in neurology.

Machine Learning- The Concept

Learning to make a diagnosis of acute stroke based on FAST (face, arm, speech, time) acronym is a form of rule-based learning. On the contrary, learning to make the same diagnosis in the setting of internuclear ophthalmoplegia and ataxia requires recall of clinical gestalt accrued from experience. ML takes the latter approach to make predictions. ML algorithm learns to generalize from a database of known examples to formulate rules for future predictions from unseen dataset. ML makes this possible by taking the concept of logistic regression a step forward by using a higher number of mathematical operations to find complex relationships in the data, e.g., risk factors and outcomes. These mathematical operations are sometimes performed in layers; each layer extracting one particular aspect of this complicated relationship. One may think of each layer representing mathematical operations akin to traditional logistic regression.[15] ML can thus be defined as the process by which an algorithm encodes statistical regularities inherent in a database of examples, into parameter weights for future predictions from new data.[16]

Terminologies used in Machine Learning

ML algorithm learns from known data to make predictions from unseen, unlabeled data. The use of data labeled by clinical experts to train ML algorithms into probabilistic and statistical models is termed *Supervised ML*.[17] Most ML algorithms of today that claim human-level efficiency has been trained via

Address for correspondence: Dr. Venugopalan Y. Vishnu,
Assistant Professor (Neurology), Room No. 704, CN Centre, Seventh
Floor, All India Institute of Medical Sciences, New Delhi, India.
E-mail: vishnuvy16@yahoo.com

Submitted: 23-Oct-2020 **Revised:** 31-Jan-2021 **Accepted:** 11-Feb-2021
Published: 18-Jun-2021

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

DOI: 10.4103/aian.AIAN_1120_20

supervised learning. A trained ML algorithm, ready to make predictions from unseen data, is referred to as a *model*.[16]

Features in ML are input variables derived from training examples. Height, weight, or pixel data of images are examples of features that may be used as input to train an ML algorithm.[15-18] *Parameters* are akin to weights in the logistic regression equations. It represents the internal values of statistical regularities inherent in the database. The ML algorithm automatically derives parameters via the learning process. When the ML algorithm is fed with successive examples in the training dataset, parameters (learnable weights) are continuously updated. Parameters dictate the accuracy of the model's predictions.[15-18] *Hyperparameter* is a configuration external to the model that is set before the model is trained and remains fixed throughout the learning process. In the model to detect papilledema, hyperparameters were set for learning rate, the batch size for processing, etc.[5] Glossary of terminologies commonly used in ML studies is summarized in Table 1.

ML algorithms go through stages of *training, validation, and external testing*. Training ML requires learnable datasets (development dataset) that are often divided into training and validation sets. Training dataset updates the parameters (learnable weights) in the model during the training process. The remaining subset of the development set (validation dataset) is used to tune the hyperparameters. External testing is done via a dataset previously unseen by the model. Datasets for development (training/validation) and external testing must remain mutually exclusive[15-18] [Figure 1].

Pertinent questions critical to understanding a paper employing machine learning tools

Was the study prospective or retrospective, observational, or randomized controlled trial? Was the protocol published a priori?

The ML algorithm to detect papilledema trained on 14,341 fundus photographs using a retrospective dataset Brain and Optic Nerve Study with Artificial Intelligence (BONSAI) and externally tested the model with 1505 fundus photographs from another retrospective dataset.[5] Of the 82 clinical AI studies reviewed in two systematic reviews and meta-analysis,

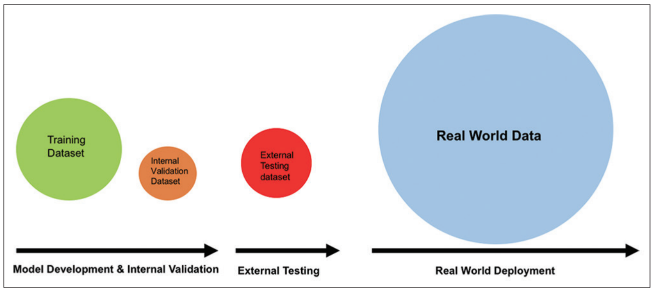


Figure 1: Model development and deployment

Table 1: Glossary of terms associated with machine learning studies

Terminology	Description
Machine learning	The process by which an algorithm encodes statistical regularities inherent in a database of examples, into parameter weights for future predictions
Supervised learning	Training a machine learning algorithm by means of previously expert labeled training examples
Model	A trained machine learning algorithm, ready to make predictions from unseen data
Features	Input variables to be used to train a machine learning algorithm
Parameters	Akin to weights in logistic regression, parameters are learnable weights representing statistical regularities in a training dataset
Hyperparameter	A configuration external to the model for e.g., learning rate and batch size for processing that is set before the model is trained and remains fixed throughout the learning process
Training	Feeding a machine learning algorithm with examples from a training dataset so that it can derive useful parameters for future predictions
Deep learning	A machine learning technique that processes information in an architecture comprising of a large number of layers, each layer extracting desired parameters incrementally from training data
Deep neural network (DNN)	A deep learning architecture with multiple layers between input and output layers
Convolutional neural network (CNN)	A class of DNN that display connectivity patterns that are analogous to that of the connectivity patterns and image processing in visual cortex
Black box	Human inability to explain the precise steps leading to the model's predictions, due to complex maze of parameters that is inscrutable to humans
Confusion matrix	Akin to contingency table in traditional studies, often used to describe performance of the model
Fine tuning (pre-initialization or warm start)	A technique in machine learning where a model is trained on an unrelated dataset of a similar data type to initialize the parameters
Transfer learning	Parameters pre-trained in the solution to one task are transferred to the new model under development to accelerate learning
Ensemble learning	A method where outputs of the two machine learning networks are combined to improve the quality of prediction and improve the overall performance of the model
F score	Generally in machine learning models there is a trade-off between recall (sensitivity) and precision obtained by varying the threshold used to categorize data into one of the two classes. In order to characterize the precision as well as recall using a single measure, the F score is used. The F score is the harmonic mean of the precision and recall which balances the contributions of these two terms. F score tasks a value between 0 and 1. For an ideal classifier with 100 percent sensitivity and specificity (or a precision and recall of 1), the F score will be 1. If either the precision or the recall is zero, F score will also be zero indicating a poor classifier

only 11 were prospective and a mere 7 were randomized controlled trials (RCTs).[19] The model to detect DR, trained on 30,000 expert labeled images from three retrospective datasets [DiaRetDB1, Kaggle (EyePACS), and Australian Tele-eye care DR], was externally tested using a prospective dataset obtained over 6 months.[6] These observational studies lay the groundwork for future RCTs. RCTs done in usual clinical care settings that compare performance metrics of expert clinicians versus ML model are important before the model can be meaningfully deployed in real-world clinical settings.[7,20] New clinical-trial guidelines specifically tailored for studies involving AI, covers framing of protocols (SPIRIT-AI Extension), and recommendation for publications of AI-based clinical trials (CONSORT-AI Extension).[21,22] *A priori* stating the study protocol helps to avoid publication bias and selective reporting of positive outcomes, which are important for mitigating the risk of distorting perceptions regarding the model's utility. Description of the research question, model training and validating strategies, outcomes, power calculation, and statistical analysis plan should be described before the commencement of the study.[18]

Why was the dataset obtained and what is its size?

If a model is being trained to determine the risk of stroke from a dataset of retrospective fundus photographs, then it is imperative to know why those fundus photographs were obtained in the first place. If the majority of those fundus photos were collected as part of diabetes or hypertensive screening, then the association of those photos with stroke increases substantially vis a vis a dataset of fundus photos obtained from a population of Leber's hereditary optic neuropathy. Depending on the dataset used for training, the trained model tasked with predicting stroke may consistently err on the side of excessive false positives or false negatives.

The ML models for detecting papilledema and DR trained on retrospective datasets are silent on the details regarding the exact indications for obtaining individual photographs in the first place.[5,6] An AI model trained to detect intracranial hemorrhage used a retrospective dataset of 46,583 non-contrast head computed tomography (CT) studies from 31,256 unique patients. Each CT scan had the indication documented in a clinical report. The details of these indications essential to interpreting the predictions of the model are, however, not mentioned in the published paper.[23]

Inclusion and exclusion criteria specified at both participant and input data level are important considerations determining generalizability of the model in real-world settings. To illustrate an example, vendor specific post-processing of retinal scan as inclusion criteria at input data level to predict DR limits generalizability of the model to centers capable of performing similar processing of retinal scans. Poor data or missing data should be reported, for example those arising out of movement artifacts while recording electrocardiograms (ECGs) or CT scans may interfere with the accuracy of labeling the ground truth by the subject experts.[21]

The size of the dataset may affect the model's capability to predict accurately from unseen data. An AI model for detection of critical findings in head CT scan showed excellent performance metrics when trained on a vast dataset comprising of 313,318 head CT scans.[24] Similarly, another model trained on a dataset of 2 million examples of labeled 12 lead ECGs, outperformed cardiology residents in detecting six types of ECG abnormalities.[25] If the dataset is small, techniques in ML namely *fine-tuning (pre-initialization or warm start)* and *transfer learning* are employed. Fine-tuning is a technique in ML where a model is trained on an unrelated dataset of a similar data type to initialize the parameters.[23] The ML model to detect papilledema was pre-initialized with a large unrelated dataset comprising of 1.28 million images over 1,000 generic objects (ImageNet), before commencing the training process.[5] In *transfer learning,* parameters pre-trained in the solution to one task are transferred to the new model under development to accelerate learning.[26] Pre-initialization with ImageNet and transfer learning with a dataset comprising of 129,450 skin lesions were employed in a model based on GoogleNet ML architecture to detect skin cancer with dermatologist level accuracy.[27]

What is the intended use of ML model in the context of the clinical pathway?

Information from various stages in patient's clinical pathway is integrated to arrive at a diagnosis which culminates in a treatment plan. It is important to know the entry point of the ML model in this clinical pathway [Figure 2]. The trained model may interact with healthcare professionals, patients, and the public. Its role in the clinical pathway may range from triage (screening patients or images) to replacing the healthcare professional for diagnosis and treatment. While generalizing the performance of the model, caution should be exercised to avoid hyperbolic inferences where a model shown to perform well in screening CT images is considered a replacement for an emergency radiologist. It is also important to know how the AI intervention was integrated into the clinical pathway at the trial site. Whether the model required vendor-specific devices,

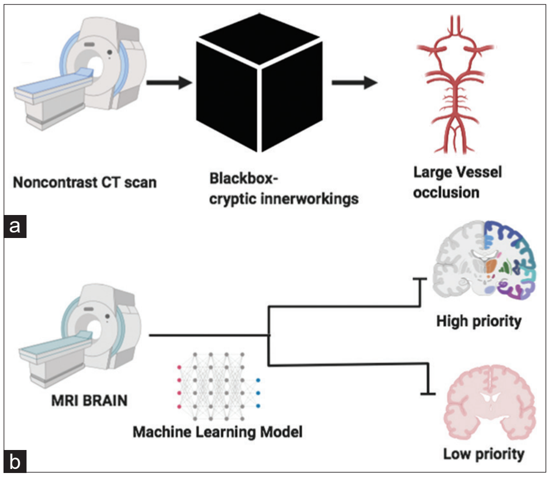


Figure 2: (a) Black box problem plagues high performing models as their inner workings are often least explainable. (b) Intended use of ML model in the context of clinical pathway should be clearly known before deploying the model at scale

support of cloud integration is a question with significant implications on its generalizability.[28,29]

Does the dataset represent the disease spectrum in the target population?

If a model to detect papilledema was trained/validated on a dataset containing only normal fundi and severe papilledema, then the model's performance to detect papilledema would be skewed. The model, if deployed in real-world scenarios, will identify severe papilledema correctly but classify a large number of mild and moderate papilledema as the normal fundus. The ML study to detect papilledema describes 2148 disks with confirmed papilledema in the training set and 360 in the external testing set without giving details about their severity grade.[5] The reader should be skeptical of the generalizability of this model's performance to a real-world scenario.

The model used to detect DR showed a high rate of false positives in the external testing phase that could be explained by the skewed representation of disease states in the development dataset, a phenomenon known as *spectrum bias*.[6] A related problem is a *class imbalance* where disease categories are not equally represented in the development dataset. This is usually addressed by balancing the classes by way of taking away instances of the overrepresented class (*undersampling*) and adding copies of the underrepresented class (*oversampling*).[18] Assigning higher weights to underrepresented class during model training is another way of balancing the class.

More extensive the gamut of development dataset, better are the chances that the model's performance shown in external testing dataset will match its performance during real-world deployment. The model for detecting DR was trained on three datasets which obtained their fundus photographs from different ethnic populations in different geographical locations. The model was externally tested in a population from Western Australia which was represented in the development dataset.[6] External testing of this model in a different ethnic population, not represented in the development dataset, may give sub-par results. [Figure 3] The pertinent questions that need asking are: does the development set represent the disease prevalence in the target population? Has the application of inclusion and exclusion criteria in obtaining the dataset caused a selection bias? Did the investigators use a sampling method, e.g., random sampling, to mitigate the risk of spectrum bias?

How was the data split between training, validation, and external testing?

The dataset that was seen by the ML during training must be kept distinct from the dataset used for external testing (unseen dataset). If possible, the two datasets should be obtained from two different populations separated in time and geographical location. The ML trained to detect papilledema achieves this goal by keeping the development (training and validation), and external testing sets separate. The training/validation set of 14,341 fundus photographs from 19 sites in 11 countries was distinct from an external testing dataset of 1505 fundus photographs obtained from five different centers in five different countries.[5]

The model trained to detect papilledema and DR split the development datasets into training and validation subsets with a ratio of 80:20.[5,6] The study to identify papilledema used five-fold cross-validation between training and validation datasets. *Five-fold cross-validation* involves running the experiment five times, each time with a random 20% sample of the development set acting as a "validation dataset". Cross-validation reduces or even eliminates the risk of selection bias.[5] Investigators training the model to detect acute neurological events derived the development and external testing datasets from the same pool of cranial CT images.[7] The mixing up of development and external testing datasets blurs the distinction of seen and unseen data by ML. The reader should bear in mind that in a real-world scenario, the model's performance trained on such mixed datasets may considerably vary from reported metrics. If a specific internal validation dataset is not mentioned as a subset of development dataset, then the reader should question whether the external testing dataset was inadvertently used for internal validation. Mixing of these datasets is a red flag in ML studies.

How was the gold standard determined?

The gold standard (data labeled by expert clinicians) against which the ML model learns to adjust its parameters is a crucial element in ML studies. If the gold standard is a clinical or paraclinical judgment, then it is essential to know the qualifications of experts arriving at the gold standard. How were the intra-rater and inter-rater variability addressed? Was it adjudicated by a panel of experts or was a majority vote used? Also, the clinicians judging the gold standard should be blinded from the ML predictions.

In the model trained to detect DR, the gold standard was determined by a single ophthalmologist.[6] The model to detect papilledema was trained on a dataset that employed a panel of neuro-ophthalmologists to determine the gold standard.[5]

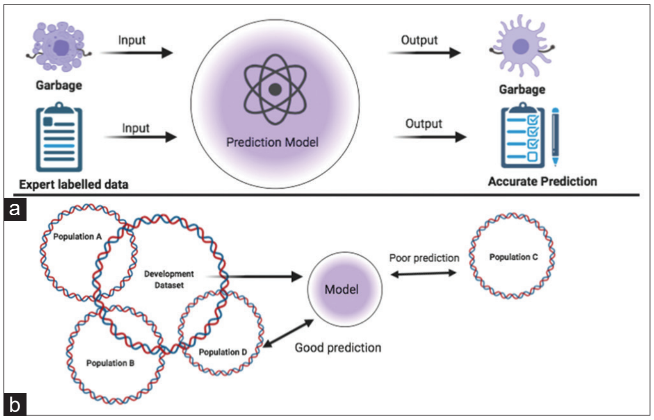


Figure 3: (a) The quality of expert labeled data in development dataset determines the models performance during external testing. (b) The model performs best in those populations that have been represented in the development dataset

In the model used to identify critical findings in head CT scan 03 senior radiologists with experience spanning 8–20 years in cranial CT interpretation determined the gold standard by mutual consensus or majority vote.[24] The quality of the gold standard used to train the model is the bedrock that determines generalizability of model's performance. [Figure 3]

What was the type of ML employed? Which version of the model was used for the study?

Early ML methods, especially those requiring processing of image-related features, used computational vision, and image processing techniques suitably adapted to the problem.[30] More recently, the deep learning-based approaches have been very successful in outperforming conventional techniques.[31] The deep learning-based techniques rely on (a) a deep neural network (DNN), so called because of their superficial resemblance to biological neural networks, acting as computational models that learn parameters in big data, (b) large data for training the models, typically comprising 50,000 or more labeled examples, (c) large computational resources typically provided by specialized graphical processing units for training the deep networks, (d) algorithmic innovations to speed up the training of deep networks.[32]

Conventional ML methods need expert intervention to predefine some features in the dataset to help the model learn, a labor intensive, and daunting task. However, the models thus learned are more open to interpretation and analysis. In contrast, the deep learning techniques often rely heavily on convolutional neural networks (CNN), a class of DNN that display connectivity patterns that are analogous to that of the connectivity patterns and image processing in visual cortex. Deep learning techniques take advantage of the hierarchical pattern in data to assemble more complex patterns using simpler and smaller patterns. The ML model to detect papilledema was trained on two CNNs (DenseNet-121 and DenseNet-201). By a method called ensemble learning, the outputs of the two networks were combined to improve the quality of prediction and improve the overall performance of the model.[5]

High performing deep learning-based models often require less prior knowledge and human effort in feature design.[5,16] The high performance of DNNs come at the cost of interpretability and robustness.[30,33] The performance metrics of ML model reported in a study is version specific. If a different version of the model was used in previous studies, then the changes brought about in the new version of the model and rationale for the changes is important considerations for generalizability.[21]

Is the reported model "continuously evolving" or "continuously learning" by design?

Most ML models are trained on development datasets to reach a certain level of efficiency. The chosen model with most satisfactory performance metrics is then "locked" in a way that it does not learn further when shown new data. In contrast, "continuously learning" models do not stop learning even beyond the stage of development. They have the ability to be continuously trained on new shown data which may cause changes in performance over time. The reported metrics of "continuously learning models" may not hold true even for the same external dataset when tested incrementally. Trials comparing "continuously learning" models with clinicians are hard to interpret and have been excluded from consideration in the newly reported CONSORT-AI extension.[21]

Does the model suffer from a black box problem?

Can the output of the model be retraced back to input on demand? The *Black Box* problem refers to human inability in explaining the precise steps leading to the ML model's predictions.[34] When an ML model is trained on massive datasets, the model may extract parameters into a complex maze of weighted connections that is inscrutable to humans. Performance and explainability in ML models are often inversely proportional. The best performing models, e.g., DNNs, are often least explainable. [Figure 2] Models with a poorer performance like linear regression and decision trees are usually most explainable. European Union General Data Protection regulation legislation requires that ML predictions be explainable, especially those that have the potential to affect users significantly. Explainable ML models instill confidence and are likely to result in faster adoption in clinical settings. There is a growing interest in interpretable models in deep learning.[35] The field of interpretable models is in the early stages of development. A deep-learning algorithm trained to detect acute intracranial hemorrhage used attention map and a prediction basis retrieved from training data in an attempt to enhance explainability.[36]

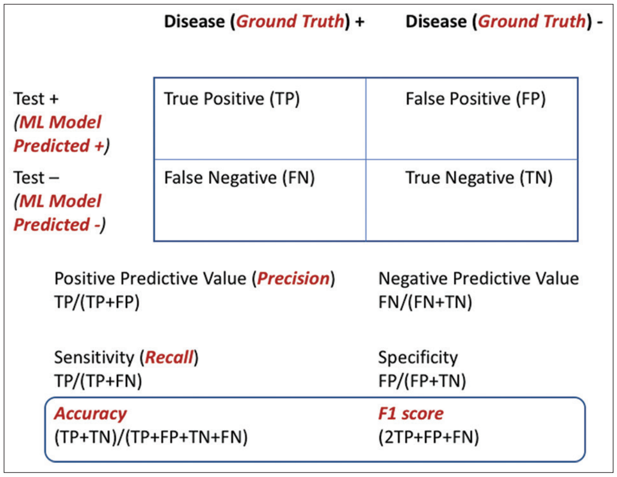
Which performance metric is being reported/optimized? How were performance errors identified and analyzed?

The output of the ML model can be a diagnostic classification or probability, a recommended action, an alarm alerting to an event or some other output. The reader should be able to differentiate performance metrics for tasks involving prediction/classification from those that show causal inference. While predictive tasks by ML model can be assessed from observational nature of big data, causal inferences usually require RCTs. To illustrate an example, the in silico study to detect papilledema that shows human level performance metrics cannot answer the counter factual scientific question "Will this model improve detection of papilledema in clinical practice?" Such counter factual predictions require RCTs performed in usual clinical care settings. The choice of suitable performance metrics is crucial to evaluate the usefulness of results.[37] For example, in the case of a rare disease classification which is likely to occur in only one out of every thousand patients screened, the ML models can give 99.9% accuracy by classifying all the examples as negative. Sensitivity and specificity may be more suitable metrics in such scenarios. ML models can make errors in output that may be hard to foresee. The reader should actively seek out reporting of such errors in published studies and the strategies used for its risk mitigation. Such errors if undetected during regulatory approval process may lead to catastrophic consequences when AI models are allowed to deploy at scale.[38]

A ML algorithm designed to recognize large vessel occlusion (LVO) patterns from CT scans without the need for contrast-enhanced imaging showed an area under the curve (AUC) for the identification of LVO as 0.87 (sensitivity: 83%, specificity: 71%, positive predictive value: 79%, negative predictive value: 76%) which improved to 0.91 when data on National Institute of Health Stroke Scale was also provided (sensitivity: 83%, specificity: 85%, positive predictive value: 88%, negative predictive value: 79%).[39] The reported performance metrics usually pertain to the model itself (F scores, Dice coefficient or AUC) or its predictions translated to relevant clinical outcomes (sensitivity, specificity, positive predictive value, negative predictive value, numbers needed to treat and AUC).[17] The names of certain performance metrics may vary from what a neurologist is used to. *Recall* reported in ML studies is equivalent to sensitivity, *precision* is equivalent to positive predictive value and *confusion matrix* is equivalent to the contingency table. Gold standard/reference test is often referred to as the *ground truth/label* in ML literature. At the minimum, papers should provide contingency table (confusion matrix), sensitivity, and specificity for easier comparisons and better understanding [Figure 4].

Is the model performance too good to be true?

Remember that the results of the ML model can only be as good as the information contained in the development set. As a corollary, the ML models should not be able to outperform an expert clinician or a panel of expert clinicians, given ample time to make a decision or diagnosis. ML models have recently revealed hitherto unseen new associations often suggesting causality between a risk factor and disease. These reports should be interpreted with caution if not borne out of RCTs. When a model predicts unexpected outputs in test datasets, the same should be confirmed in different patient cohorts to mitigate the effects of confounding factors, artifacts or flaws in study designs.



	Disease (*Ground Truth*) +	Disease (*Ground Truth*) -
Test + (*ML Model Predicted +*)	True Positive (TP)	False Positive (FP)
Test – (*ML Model Predicted -*)	False Negative (FN)	True Negative (TN)

Positive Predictive Value (*Precision*)
TP/(TP+FP)

Negative Predictive Value
FN/(FN+TN)

Sensitivity (*Recall*)
TP/(TP+FN)

Specificity
FP/(FP+TN)

Accuracy
(TP+TN)/(TP+FP+TN+FN)

F1 score
(2TP+FP+FN)

Figure 4: Comparison of traditional performance metrics with that of machine learning methods (terms in italics are machine learning terminology)

A model can learn to *overfit* to development dataset by learning patterns that are too specific to the dataset. For example, a model learnt to classify a skin lesion as malignant if the image had a ruler in it. The model had learnt to recognize the spurious signal of the presence of a ruler in the images of development dataset with increased chances of a cancerous lesion.[27] Similarly, another model learnt to diagnose pneumonia from chest X-rays based on the increased association between a portable X-ray machine used for recording X rays and pneumonia.[40] These are examples of the model adjusting its parameters to spurious signals within the development dataset. The reader must suspect overfitting when there are vast differences in performance metrics of the model between validation and external testing.

Is the study repeatable and reproducible? Are source code and datasets available for scrutiny?

Identical images shown to the model at separate times should yield identical predictions. A repeat imaging may contain the minor difference of pixels but should give similar predictions from the model (*repeatability*). When similar images are taken with different hardware at different institutions employing different operators and protocols, the differences in predictions should be quantified (*reproducibility*).

Reproducibility in ML has been a critical challenge that has become even more important in the context of medical applications.[41,42] The reproducibility may be formalized at the following three levels[43]:

(a) Model reproducibility — whether the description of the model is detailed enough such that another researcher can independently write the code for the model and reproduce the claimed results.
(b) Model+code reproducibility — The source code that led to the claimed performance along with the model descriptions may be shared with researches who can independently run the code on their data without having to worry about the grueling details of the model.
(c) Model+code+data reproducibility — Many ML labs also release the data along with the model and code for others researches to replicate, explore, and experiment.

In practice, the “Model reproducibility” is almost impossible to achieve especially for the latest high-performing ML models because of the many little details usually left unspecified in the model descriptions which may significantly impact the performance of the models. Reporting of all the three aspects of the tool namely model, code, and development data generally leads to reproducible results when applied to other datasets. Sharing of raw clinical data is often not a pragmatic solution due to institutional patient-privacy policies. Even when the raw clinical development data are available, reproducibility is not guaranteed as the models may be non-deterministic due to parallel processing (different runs may give different results) or use specific random numbers seeds.

Does the AI intervention affect patient outcomes?

The outcomes of standalone in silico experiments that contrast with the messy world of usual clinical practice should be interpreted with caution. While important for laying the groundwork for future clinical trials, they themselves do not show causal inference pertaining to clinical outcomes. "Whether the reported model will affect clinical outcomes?" is a counter factual question implying causal inference that needs RCTs for a conclusive answer, a rare phenomenon in today's world of so called "AI clinical trials."[19] An RCT enrolled 350 pediatric patients where cataract assessment with or without an AI platform was compared to diagnose cataract and provide treatment recommendation (surgery or follow up). The AI reported metrics for diagnostic accuracy and treatment recommendations were 87% (sensitivity 90%, specificity 86%) and 71% (sensitivity 87%, specificity 44%), respectively. The consultants performed significantly better by comparison for accuracy of diagnosis (99%, sensitivity 98%, specificity 99.6%) and treatment recommendation (97%, sensitivity 95%, specificity 100%) ($P < 0.001$ for both). The same AI model had earlier shown a significantly higher performance in a non-randomized clinical trial setting for accuracy of diagnosis and treatment recommendation (98% and 93%, respectively).[20] When a clinician is evaluating a cataract or a skin lesion they are not merely analyzing a photograph in isolation but performing a holistic assessment in the context of patient's history and physical examination. A study compared a DNN with 21 board certified dermatologists in diagnosing skin cancer based on analysis of photographs of skin lesion shown to both groups. The study reported dermatologist level accuracy in diagnosing skin cancer by DNN.[27] Drawing inferences from such comparisons that are far removed from usual clinical practice is not helpful in answering the above stated counterfactual question. A new tool is worth its salt only if it can make a dent in the clinical outcomes.

Conclusion

ML often championed as a solution to prediction problems from big data, also evoke concerns that artificial intelligence in clinical medicine is overhyped and requires proper guidance, knowledge, or expertise, to mitigate methodological shortcomings, poor transparency, and poor reproducibility. ML models learning from expert labeled data predict with consistency, speed, and lack of fatigue, a feat rarely achievable by humans. Given an abundance of time and information, the expert clinician should be able to deliver comparable predictions, a useful benchmark while evaluating the performance of ML models. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement published in 2015 provides guidance on key items to report when describing studies on developing, evaluating (or validating), or updating clinical prediction models. A new initiative to develop a version of the TRIPOD statement (TRIPODML) focusing on ML prediction algorithms is underway.[44] The process of training/validation and external testing should not be restricted to a one-time event but should be an ongoing process spanning different geographical locations and ethnic populations. Predictive performance metrics of ML models should not be confused with causal inference between its input and output. Prospective RCTs comparing the output of the ML model and clinicians are necessary to unravel its true utility as a clinical tool. Unbelievable performance metrics displayed by ML model should raise a red flag and be investigated further. A useful checklist to critically appraise a ML paper is summarized in Table 2. ML and the clinical gestalt must compete in a RCT before they can complement each other in a real-world deployment to improve diagnosis, prognosis, monitoring, and administration of treatments, in realizing the common aim of improving health outcomes.

Disclosures: None

Search strategy and selection criteria

We searched PubMed, Medline, and Google scholar for relevant articles published in English between Jan 1, 2010 and Sep 30, 2020 using the terms "artificial intelligence," "ML," "supervised learning," "deep learning," "deep neural network," "convolutional neural network," "big data," "fine tuning," "black box," "critical appraisal," "evidence-based medicine." We selected studies that were relevant to the field of medicine and neurology. We also searched the reference lists of retrieved articles. We then selected the most relevant references, paying particular attention to studies within the past 5 years and studies with large samples, control groups, and reduced bias. We retained some older studies for their importance.

Table 2: Summary of key questions in critical appraisal of a machine learning research paper

S.No	Questions
1	Was the study prospective or retrospective, observational, or randomized controlled trial?
2	Was the protocol published *a priori*?
3	Why was the dataset obtained, and what is its size?
4	What is the intended use of ML model in the context of the clinical pathway?
5	Does the dataset represent the disease spectrum in the target population?
6	How was the data split between training, validation and external testing?
7	How was the gold standard determined?
8	What was the type of ML employed? Which version of the model was used for the study?
9	Is the reported model "continuously evolving" or "continuously learning" by design?
10	Does the model suffer from a black box problem?
11	Which performance metric is being reported/optimized? How were performance errors identified and analyzed?
12	Is the model performance too good to be true?
13	Is the study repeatable and reproducible? Are source code and datasets available for scrutiny?
14	Does the AI intervention affect patient outcomes?

Acknowledgements

The authors acknowledge Biorender for creation of Figures 2 and 3.

Ethical approval

Not applicable

Financial support and sponsorship

Supported by Grant from Department of Biotechnology, Government of India (BT/PR33179/AI/133/16/2019).

Conflicts of interest

There are no conflicts of interest.

References

1. Leslie-Mazwi TM, Lev MH. Towards artificial intelligence for clinical stroke care. Nat Rev Neurol 2020;16:5-6.
2. Pedersen M, Verspoor K, Jenkinson M, Law M, Abbott DF, Jackson GD. Artificial intelligence for clinical decision support in neurology. Brain Commun 2020;2:fcaa096. doi: 10.1093/braincomms/fcaa096.
3. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, *et al*. Artificial intelligence in healthcare: Past, present and future. Stroke Vasc Neurol 2017;2:230-43.
4. Obermeyer Z, Emanuel EJ. Predicting the future-Big data, machine learning, and clinical medicine. N Engl J Med 2016;375:1216-9.
5. Kohane I. AI for the eye-Automated assistance for clinicians screening for papilledema. N Engl J Med 2020;382:1760-1.
6. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney M, Mehrotra A. Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care. JAMA Netw Open 2018;1:e182665.
7. Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, *et al*. Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat Med 2018;24:1337-41.
8. Gupta U, Bansal H, Joshi D. An improved sex-specific and age-dependent classification model for Parkinson's diagnosis using handwriting measurement. Comput Methods Programs Biomed 2020;189:105305. doi: 10.1016/j.cmpb.2019.105305.
9. Khajuria A, Joshi P, Joshi D. Comprehensive statistical analysis of gait parameters in neurodegenerative diseases. J Neurophysiol 2018;50:38-51.
10. Gupta K, Khajuria A, Chatterjee N, Joshi P, Joshi D. Rule based classification of neurodegenerative diseases using data driven gait features. Health Technol 2018. doi: 10.1007/s12553-018-0274-y.
11. Joshi D, Khajuria A, Joshi P. An Automatic Non-Invasive Method for Parkinson's disease Classification. Comput Methods Programs Biomed 2017;145:135-45. doi: 10.1016/j.cmpb.2017.04.007.
12. Daoud H, Bayoumi MA. Efficient epileptic seizure prediction based on deep learning. IEEE Trans Biomed Circuits Syst 2019;13:804-13.
13. Claassen J, Doyle K, Matory A, Couch C, Burger KM, Velazquez A, *et al*. Detection of brain activation in unresponsive patients with acute brain injury. N Engl J Med 2019;380:2497-505.
14. Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J; Alzheimer's Disease Neuroimaging Initiative. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 2015;104:398-412.
15. Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: Users' guides to the medical literature. JAMA 2019;322:1806-16.
16. Zador AM. A critique of pure learning and what artificial neural networks can learn from animal brains. Nat Commun 2019;10:3770.
17. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, *et al*. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nat Med 2020;26:1320–4.
18. Faes L, Liu X, Wagner SK, Fu DJ, Balaskas K, Sim DA, *et al*. A clinician's guide to artificial intelligence: How to critically appraise machine learning studies. Transl Vis Sci Technol 2020;9:7.
19. Topol EJ. Welcoming new guidelines for AI clinical research. Nat Med 2020;26:1318–20.
20. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, *et al*. Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. EClinicalMedicine 2019;9:52-9.
21. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI Extension. BMJ 2020;370:m3164. doi: 10.1136/bmj.m3164.
22. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group; SPIRIT-AI and CONSORT-AI Steering Group; SPIRIT-AI and CONSORT-AI Consensus Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. Nat Med 2020;26:1351-63.
23. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, Suever JD, Geise BD, Patel AA, *et al*. Advanced machine learning in action: Identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. NPJ Digit Med 2018;1:9.
24. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, *et al.* Deep learning algorithms for detection of critical findings in head CT scans: A retrospective study. Lancet 2018;392:2388-96.
25. Ribeiro AH, Ribeiro MH, Paixao GM, Oliveira DM, Gomes PR, Canazart JA, *et al*. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 2020;11:1760.
26. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data 2016;3:9. doi: 10.1186/s40537-016-0043-6.
27. Esteva A, Kuprel B, Novoa R, Ko J, Swetter SM, Blau HM, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
28. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17:195.
29. Pooch EH, Ballester PL, Barros RC. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. arXiv 2019. http://arxiv.org/abs/19090.01940.
30. Harry W. Computational Vision. Elsevier; 2016.
31. Yann L, Bengio Y, Hinton G. Deep Learning. Nature 2015;521:436–44.
32. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform 2016;7:29. doi: 10.4103/2153-3539.186902.
33. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E, *et al*. Deep learning for computer vision: A brief review. Comput Intell Neurosci 2018;2018:7068349. doi: 10.1155/2018/7068349.
34. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1**:**206–15.
35. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv: 1702.08608 (2017).
36. Lee H, Yune S, Mansouri M, Kim M, Tajmir SH, Guerrier CE, *et al*. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nat Biomed Eng 2019;3:173-82.
37. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In: Sattar A, Kang B, editors. AI 2006: Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science. Vol 4304. Berlin, Heidelberg: Springer. doi: 10.1007/11941439_114.
38. Oakden-Rayner L, Dunnmon J, Carniero G, Re C. Hidden Stratification causes clinically meaningful failures in machine learning for medical imaging. arXiv http://arxiv.org/abs/1909.12475 (2019).
39. Olive-Gadea CM, Crespo C, Granes C, Hernandez-Perez M, de la Ossa NP, Laredo C, *et al*. Deep learning based software to identify large vessel occlusion on noncontrast computed tomography. Stroke 2020;51:3133-7.
40. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15:e1002683.

41. Beam AL., Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. JAMA 2020;323:305-6.
42. McDermott MB, Wang S, Marinsek N, Ranganath R, Ghassemi M, Foschini L, *et al*. Reproducibility in machine learning for health. arXiv preprint arXiv: 1907.01463 (2019).
43. Tatman R, VanderPlas J, Dane S. A practical taxonomy of reproducibility for machine learning research. 2nd Reproducibility in Machine Learning Workshop at ICML 2018, Stockholm, Sweden.
44. Collin GS. Moons KG. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577-9.