Viewpoint

Key Technology Considerations in Developing and Deploying Machine Learning Models in Clinical Radiology Practice

Viraj Kulkarni[1*], MSc; Manish Gawali[1*], BE; Amit Kharat[1,2*], DPhil

[1] DeepTek Inc, Pune, India

[2] D Y Patil University, Pune, India

[*] all authors contributed equally

Corresponding Author:
Manish Gawali, BE
DeepTek Inc
2rd Floor, Alacrity Innovation Centre
3, Baner Rd, Pallod Farms, Baner
Pune, 411045
India
Phone: 91 72760 60080
Email: manish.gawali@deeptek.ai

Abstract

The use of machine learning to develop intelligent software tools for the interpretation of radiology images has gained widespread attention in recent years. The development, deployment, and eventual adoption of these models in clinical practice, however, remains fraught with challenges. In this paper, we propose a list of key considerations that machine learning researchers must recognize and address to make their models accurate, robust, and usable in practice. We discuss insufficient training data, decentralized data sets, high cost of annotations, ambiguous ground truth, imbalance in class representation, asymmetric misclassification costs, relevant performance metrics, generalization of models to unseen data sets, model decay, adversarial attacks, explainability, fairness and bias, and clinical validation. We describe each consideration and identify the techniques used to address it. Although these techniques have been discussed in prior research, by freshly examining them in the context of medical imaging and compiling them in the form of a laundry list, we hope to make them more accessible to researchers, software developers, radiologists, and other stakeholders.

(JMIR Med Inform 2021;9(9):e28776) doi: 10.2196/28776

KEYWORDS

artificial intelligence; AI; machine learning; deep learning; radiology; privacy; neural networks; deployment

Introduction

Although radiology imaging has emerged as an indispensable tool in diagnostic medicine, there is a worldwide shortage of qualified radiologists to read, interpret, and report these images [1,2]. The volume of images is growing faster than the number of radiologists. The high workload that this causes leads to errors in diagnosis because of human fatigue, unacceptable delays in reporting, and stress and burnout in radiologists. On the other hand, artificial intelligence (AI) and machine learning models have shown remarkable performance in the automated evaluation of medical images [3-5]. In this situation, hospitals are increasingly drawn toward adopting computer-aided detection technologies for processing scans. These technologies show considerable promise in improving diagnostic accuracy, reducing reporting time, and boosting radiologist productivity. Supervised machine learning, the most common form of machine learning, works in two phases. In the first phase, the algorithm implemented as a software reads a training data set consisting of images along with their corresponding labels. It processes these data, extracts patterns from it, and learns a function that maps an input image to its corresponding label. The learned mapping function along with the extracted patterns are mathematically represented in the form of the trained model. This is called the *training* phase. In the second phase, called the *inference* phase, the trained model is used to read input images and make predictions. Artificial neural networks are a class of machine learning algorithms; artificial neural networks with many layers are called deep neural networks. In the literature, the terms deep learning, AI, and artificial neural networks tend to be used interchangeably. In this paper, we use *machine learning* to broadly refer to all the terms mentioned earlier in addition to conventional machine learning algorithms, such as linear regression, support vector machines, decision trees, and random forests.

The development of machine learning models for radiology involves many challenges. High-quality training data are vital for good model performance [6] but are difficult to obtain. Available data may lack volume or diversity. It may be scattered across multiple hospitals. Even if the image data are available, they may not be labeled. Radiology scans suffer from a high degree of interreader variability, where 2 or more radiologists label the data inconsistently [7,8]; this may lead to noise or uncertainty in the ground truth labels. The distribution of target classes may be heavily skewed, especially for rare pathologies. This imbalance in class representation is often accompanied by unequal misclassification costs across classes. Care must be taken when dealing with imbalanced data sets, and this sometimes requires using special performance measures [9]. A model that works well on data from one hospital may perform poorly on data from a different hospital [10]. Similarly, a model deployed in practice at a hospital may experience a gradual decay in performance at the same hospital [11]. Machine learning models have been shown to be vulnerable to malicious exploits and attacks [12-14]. To support adoption by radiologists, the deployed models should be able to explain their decisions [15], and they should not discriminate patients on the basis of gender, ethnicity, age, income, among others [16].

This study has a simple structure. In the Key Considerations section, we enumerate the key considerations that machine learning researchers should acknowledge and address. For each consideration, we describe the common challenges and their significance before suggesting solutions to overcome them. In the Conclusions section, we discuss other overarching limitations that hinder the adoption of machine learning in clinical radiology practice.

Key Considerations

Insufficient Training Data

Machine learning models are data hungry, and their performance depends heavily on the characteristics of the data used to train them [6]. The training set size has a direct and significant effect on the performance of the models. However, the heterogeneity and diversity of the training data influence the ability of the models to generalize to unseen data sources [17]. To develop robust machine learning models, researchers need access to large medical data sets that adequately represent data diversity in terms of population features such as age, gender, ethnicity, and medical conditions and imaging features such as equipment manufacturers, image capture settings, and patient posture. Most available data sets in medical imaging do not meet these requirements [18-20]. As many critical conditions have low rates of occurrence, very little data are available for them. Machine learning models trained using these scanty data to diagnose rare conditions fail to perform well in practice even if they demonstrate good performance in retrospective evaluations.

Several methods have been proposed for dealing with insufficient data for training models. Data augmentation techniques including geometric transformations and color-space transformations can enhance the quantity and variety of training data [21]. Generative adversarial networks have shown success in generating synthetic images for rare pathologies, which can be further used for model training [22]. Although these techniques allow models to be trained on scarce data by artificially increasing the variation in the data set, they cannot serve as a substitute for high-quality data.

Decentralized Data Sets

Many medical data sets are naturally distributed across multiple storage devices connected to networks owned by different institutions. In traditional machine learning settings, these data sets need to be consolidated into a single repository before training the models. Moving large volumes of data across networks poses several logistical and legal challenges [23]. Government policies such as the General Data Protection Regulation [24], the Health Insurance Portability and Accountability Act [25], and the Singapore Personal Data Protection Act [26] also stipulate restrictions on sharing and movement of data across national borders.

Privacy-preserving distributed learning techniques such as federated learning [27] and split learning [28] enable machine learning models to train on decentralized data sets at multiple client sites without moving the data and compromising privacy. Implementing these techniques, however, entails additional overheads, which may render the exercise unfeasible. These overheads include the high cost of developing software that supports these technologies, the high network communication bandwidth required, the orchestration effort in deploying it at multiple sites, and the possibly reduced performance of the predictive models [29]. Federated learning generates a global shared model for all clients, leading to situations where, for some clients, the local models trained on their private data perform better than the global shared model. In such situations, additional personalization techniques may be required to fine-tune the global model individually for each client [30].

High Cost of Annotations

Supervised machine learning requires the annotation of radiology images before they can be used to train the model. Image-level annotations classify each image into one or more classes, whereas region-level annotations highlight regions within an image and classify each region into one or more classes. As the predictive performance of the model is directly influenced by the quality of the annotations, it is imperative that the data are annotated by qualified radiologists or medical practitioners [31]. This makes the process of annotation exorbitantly expensive in many cases.

Several efforts have been made to use natural language processing (NLP) techniques to automatically annotate images by extracting labels from radiology text reports [32-34]. Semisupervised approaches can be used when a small amount of labeled data are available along with a larger amount of unlabeled data [35,36]. As manual annotations are expensive, AI-based automated image annotation techniques can be considered [37].

Ambiguous Ground Truth

As hospital data sets usually contain images accompanied by their text reports, many projects are kickstarted by using NLP techniques to automatically annotate the images using the reports. Radiology reports, however, vary widely in their comprehensiveness, style, language, and format [38]. Even if state-of-the-art NLP manages to accurately extract all the findings from the text report, the report itself may not mention all the findings. Olatunji et al [39] showed that there is a large discrepancy between what radiologists see in an image and what they mention in the report; reporting radiologists usually document only those findings that are relevant to the immediate clinical context and are likely to miss reporting nonactionable or borderline findings.

Radiology images suffer from significant interreader variability, where 2 or more experts may disagree on the findings from a scan [7,8,40-43]. Sakurada et al [44], for instance, report low interreader κ values ranging from 0.24 to 0.63 for assessment of different pathologies from chest radiographs. In practice, annotation workflows generally engage a single reader to assign ground truth labels to images. An improvement over this involves engaging multiple independent readers and considering their majority vote as the ground truth label. However, single reader or majority vote approaches may miss labeling challenging but critical findings.

This risk can be mitigated by using multiphase reviews [45] or expert adjudication [46] to create high-quality labels. Majkowska et al [46] showed that adjudication improved the consensus among radiologists to 96.8% compared with 41.8% after the first independent readings when assessing chest radiographs. Raykar et al [47] proposed a probabilistic approach to determine the *hidden* ground truth from labels assigned by multiple radiologists and demonstrated that this method is superior to majority voting. In some clinical settings, radiology imaging is used for initial screening before conducting subsequent confirmatory tests. For example, chest x-ray scans may be used as a first-line test before subsequently conducting a computed tomography scan, laboratory tests, or biopsy. Data from these subsequent tests, if available, should be used to validate and correct the labels assigned to the images from the screening test. In situations where human-labeled ground truth is noisy or ambiguous, developing a process to reduce variability and improve label quality may yield better models than attempts to improve model performance on the original labels by other means.

Imbalance in Class Representation

Class imbalance occurs when all label classes are not equally represented in the training data set [48]. This is a common situation when building binary classifiers for medical data sets where the number of *normal* examples in which the target abnormality is absent is many times larger than the number of *abnormal* examples in which it is present. As machine learning models are usually trained by optimizing a loss function across all training examples, the trained models tend to favor the majority class over the minority class. Researchers have empirically evaluated the adverse effect of class imbalance on classification performance in several studies [9,49-53].

Class imbalance can be handled at the data level or the algorithmic level. Resampling strategies can be used to address imbalances in the training data by either undersampling the majority classes or oversampling the minority classes. Many comparative evaluations of these approaches exist, sometimes with contradictory conclusions. Drummond et al [54], for instance, argued that undersampling works better than oversampling, whereas Batista et al [55] reported superior performance using oversampling. However, we caution the reader against hasty generalizations and note that these comparisons are highly dependent on the data set, the machine learning algorithm, the sampling technique used, and the parameters of the experiments. Chawla et al [56] proposed the synthetic minority oversampling technique, a technique to generate synthetic examples to balance the data set, and showed that the combination of synthetic minority oversampling technique and undersampling performs better than plain undersampling or oversampling. Similarly, oversampling can also be performed using geometric augmentations, color-space augmentations, or generative models to produce synthetic images. An imbalance in the number of examples can also be addressed at the algorithmic level using methods such as one-class classification, outlier or anomaly detection, regularized ensembles, and custom loss functions [9,57-60].

Asymmetric Misclassification Costs

Standard machine learning settings assume that all misclassifications between classes are equal and incur the same penalty. This assumption is not true for many medical imaging problems. For example, the cost of classifying a *normal* scan as *abnormal* may be very different from the cost of classifying an *abnormal* scan as *normal*.

This asymmetrical nature of the classification problem can be handled either at the time of deployment or during development. The trained model can be tuned to achieve higher sensitivity or specificity according to the requirements at deployment time. Alternatively, the variation in misclassification penalties can be represented as a cost matrix, where each element C(i,j) represents the penalty of misclassifying an example of class i as class j. The model can then be trained by minimizing the overall cost as defined by the asymmetrical loss function. For more details, we refer the reader to the literature on cost-sensitive learning [9,61,62].

Relevant Performance Measures

Machine learning researchers and practitioners tend to ignore the question of how model performance should be evaluated in cases of imbalanced data sets and asymmetric misclassification costs. Most binary classification models produce a continuous-valued output score. This score is converted into discrete binary labels using a cutoff threshold. Owing to its simplicity, it is tempting to use accuracy, defined as the percentage of predictions that are correct, as a measure of performance. However, in the case of imbalanced data sets, accuracy is ineffective and provides an incomplete and often misleading picture of the ability of the classifier to discriminate between the two classes [63,64].

Using two or more measures such as sensitivity, specificity, and precision, provides a better picture of the discriminative performance of a classifier [65]. However, these measures depend on the cutoff threshold mentioned earlier. Furthermore, the decision to set the threshold is often guided not by technology but by business or domain concerns. Comparing the two models by considering multiple performance measures across different operating thresholds is challenging. The receiver operating characteristic curve, on the other hand, captures the model performance at all threshold operating points. The area under the receiver operating characteristic curve (AUROC) thus serves as a single numerical score that represents the performance of the model across all operating threshold points. This has made AUROC a metric of choice for reporting the classification performance of machine learning models. Unfortunately, AUROC too can be deceptive when dealing with imbalanced data sets and may provide an overly optimistic view of performance [9]. The precision-recall curve and the area under it are more suitable for describing classification performance when data sets are imbalanced [66,67]. Drummond and Holte proposed cost curves that describe the classification performance over asymmetric misclassification costs and class distributions [68,69]. Table 1 shows how accuracy can be misleading because of imbalanced data sets.

Table 1. Example illustrating how accuracy can be misleading in case of imbalanced data sets.

	Predicted as negative	Predicted as positive	Total
Actual negative	80	10	90
Actual positive	5	5	10
Total	85	15	100

In the confusion matrix mentioned earlier, of 100 test examples, 90 are negative and 10 are positive. The classifier predicts 85 of them as negative and 15 as positive. This gives a high accuracy of 0.85 and a high specificity of 0.89. However, the complete picture is seen when we consider the low sensitivity of 0.50 and precision of 0.33.

Generalization of Models to Unseen Data Sets

Machine learning models are routinely evaluated on a hold-out set taken from the same source as the training set [70]. The available data are divided into two parts. One part is used to train and validate the models. The second part, called the test set or hold-out set, is used to estimate the final performance of the trained model when deployed. The underlying premise is that the data used to train the model are representative of the data that the model will encounter during clinical use. This assumption is often violated in practice, and this makes the performance on the hold-out set an unreliable indicator of future performance in clinical deployment.

Poor generalization of models to diverse patient groups is one of the biggest hurdles for the adoption of AI and machine learning in health care. One reason for the poor generalization is the difference in the image characteristics between images from the training sites and those from the deployment site. This variation, also known as data set shift, can occur because of differences in hospital procedures, equipment manufacturers, image acquisition parameters, disease manifestations, patient populations, among others. Owing to the data set shift, models trained using data from one hospital may perform poorly on data from another hospital [71]. We note here that this inability to generalize to data sets from an unseen origin is different from the problem of overfitting, where the model shows poor performance even on test sets from the same origin. Learning irrelevant confounders instead of relevant features is another reason why models fail to generalize to data from unseen origins. Machine learning models are notorious for exploiting confounders in the training data. For example, Zech et al [72] showed that a pneumonia classification model trained on data from 2 hospitals learned to leverage the difference between prevalence rates at the 2 hospitals instead of the relevant visual features.

Data augmentation can improve model generalization by increasing the variations in the training set [73]. Image processing techniques, including standardization, normalization, reorientation, registration, and histogram matching, can be used to harmonize images sourced from different origins and remove domain bias. However, Glocker et al [74] showed that even with a state-of-the-art image preprocessing pipeline, these techniques for harmonization were unable to remove scanner-specific bias, and machine learning models were easily able to discriminate between the different origins of the data.

Domain adaptation techniques can be used to fine-tune models to a new target domain by narrowing the gap between the source and target domains in a domain-invariant feature space [75-79]. On the other hand, domain generalization techniques attempt to train models that are sensitive only to features relevant to the classification task but insensitive to confounding features that differentiate between the domains [80-85].

Model Decay

Model decay refers to the phenomenon in which the performance of a deployed machine learning model deteriorates over time [11]. Supervised machine learning algorithms extract patterns from the training data to learn a mapping between independent input variables and a dependent target variable. This process involves making an implicit assumption that the data encountered in deployment will be stationary and will not change over time; this assumption is often violated in practice because of the changes in hospital workflows, imaging equipment, patient groups, evolving adoption of AI solutions, among others.

Model decay occurs owing to changes in the underlying data. These changes can be broadly classified into three types: (1)

covariate shift occurs when there are changes in the distribution of the independent input variables (eg, the average age of the population increases over time); (2) *prior probability shift* occurs when there are changes in the distribution of the dependent target variables (eg, the prevalence of a particular disease in the target population may change because of seasonality or an epidemic); and (3) *concept drift* occurs when there are changes in the relationship between the independent and dependent variables (eg, changes in a hospital's diagnostic protocols or a radiologist's interpretation regarding which visual manifestations should or should not be considered indicative of a pathology). These changes can be sudden, gradual, or cyclic.

Detecting model decay requires continuous monitoring of the deployment time performance against a human-labeled subsample of the data. If the performance drops below a predetermined threshold, an alarm is triggered, and the model is retrained or fine-tuned using the most recent data. This retraining can also be conducted periodically as a routine maintenance activity. For more details, including theoretical frameworks for understanding model decay or practical solutions, readers can refer to additional reviews [11,86-89].

Adversarial Attacks

An adversarial example is constructed by deliberately injecting perturbations in the original image to trick the model into misclassifying the label for that image [12]. Machine learning models are susceptible to manipulation using such adversarial examples [90,91]. Data poisoning attacks [13] introduced adversarial examples in the training data to manipulate the diagnosis of the model being developed. On the other hand, evasion attacks [14] use adversarial examples to influence predictions during deployment. Health care is a huge economy, and many decisions regarding diagnosis, reimbursements, and insurance may be governed or assisted by algorithms in the near future. Hence, the discovery of these vulnerabilities has raised pressing concerns regarding the safety and usability of machine learning models in clinical practice.

Qayyum et al [92] provided a detailed taxonomy of defensive techniques against adversarial attacks by grouping them into three broad categories: (1) reconstructing the training or testing data to make it more difficult to manipulate [90,93-96], (2) modifying the model to make it more resilient to adversarial examples [97-101], and (3) using auxiliary models or ensembles to detect and neutralize adversarial examples [102-106]. Adversarial attacks and their countermeasures are an evolving research area, and there are excellent reviews for the same [107-109].

Explainability

The power of neural networks to uncover hidden relationships between variables and use them to make predictions is tempered by one disadvantage: the exact process the neural network uses to arrive at a decision is unclear to humans. This is why neural networks are sometimes called black boxes whose inner workings cannot be observed. To what extent can we delegate decision-making to machines while we remain unaware of how the machine arrives at a decision is a key question that stands in the way of adopting algorithms in many industries, including autonomous vehicles, law, finance, among others.

Algorithmic explainability is especially important in medicine, where stakes are high, and the field is prone to litigation. In the context of radiology, explainability can be improved by using localization models that highlight the region of interest within the scan that is suspected to contain the abnormality instead of classification models that only indicate the presence or absence of an abnormality. However, the development of localization models also requires training data to have region-based annotations in the form of bounding boxes or free-form masks. Where region-based annotations are not available, saliency maps [110] and explainability frameworks [111] can be used to identify a region within the image that most contributes to a particular decision. Another way to improve the user's trust in the models is to predict a confidence score in addition to the prediction. For example, instead of merely stating the prediction "Probability of Tuberculosis: 75%," the system should also state the model's confidence "Probability of Tuberculosis: 75%, Confidence in this prediction: Low." Deployment settings where predictive models are used to autonomously make decisions demand more stringent conditions of explainability than settings where the models are used to guide humans who make the final decisions. A comprehensive analysis of explainers in the domain of computer vision was performed by Buhrmester et al [112].

There have been calls to limit the use of AI and machine learning only to rule-based systems in fields where algorithmic decisions affect human lives [113]. These systems are transparent and can trace the relationship between the input and the output as a sequence of rules that humans can understand. We find two problems with this approach. First, one of the chief advantages of using neural networks is that they can model complex relationships that *humans cannot understand*, and this is precisely what makes them so effective. Second, making decision systems transparent and explainable also makes them vulnerable to malicious attacks. A transparent rule-based method to make decisions can be *hacked*, *gamed*, or exploited more easily than a black box system [114,115].

Fairness and Bias

Algorithmic systems play a key role in guiding decisions that impact the delivery of health care to patients. Therefore, it is desirable that these systems are free of societal biases and their decisions are fair and equitable. Unfortunately, many existing data sets [18,43] reflect the biases of the societies that they represent [116], and it is difficult to detect and remove bias inherent in the training data. Obermeyer et al [16] showed, for example, that a widely used algorithmic system exhibited racial bias against Black patients, which reduced the number of Black patients eligible for extra care by more than half.

In principle, a predictive model is considered fair if it does not discriminate patients on the basis of sensitive variables such as gender, ethnicity, disability, and income. However, translating this seemingly simple principle into practice is a challenging issue. Researchers have developed numerous mathematical definitions of fairness and techniques to implement them [117]. One technique, for example, excludes sensitive variables from the input when training the model. Another technique is to tune the model so that it demonstrates the same level of performance as measured by sensitivity, specificity, among others, across all groups defined by the sensitive variables. Corbett-Devies and Goel [118] show that although appealing, these techniques suffer from significant statistical limitations and may adversely affect the same groups they were designed to protect. Pleiss et al [119] show how different definitions of fairness can be mutually incompatible, and a model designed to comply with one definition may violate another equally valid definition.

Algorithmic bias and fairness are evolving fields of research that lie at the intersection of machine learning, public policy, law, and ethics. We believe that fairness is not inherently a technological problem but a societal one. Coercing technology to solve it can lead to automated systems that tick the right boxes for some arbitrary definition of fairness but eventually end up worsening social inequality and discrimination behind a veneer of technical neutrality [120].

Clinical Validation

A comprehensive evaluation to assess the predictive performance and clinical utility of a model must be conducted before it can be deployed in clinical practice. When a model is evaluated on a hold-out set collected from the same sources from which the training data are collected, the evaluation is called an *internal* validation. When a data set from an *unseen* source is used to evaluate the model, the evaluation is called *external* validation. As described earlier in the section *Generalization of Models to Unseen Data Sets*, the lack of generalization to unseen data sources is one of the biggest challenges in the adoption of machine learning in practice. Despite this, only a fraction of the published studies report the results of an external validation [121]. Mahajan et al [122] presented examples to advocate the case for independent external validation of models before deployment and described a framework for the same. Park et al [31] proposed a methodology with a checklist for evaluating the clinical performance of the models. The TRIPOD statement [123] provides guidelines for transparent reporting of the development and validation of prediction models for prognosis and diagnosis models. Although retrospective evaluations allow machine learning developers to test their models on large and diverse data sets, prospective evaluations allow testing in real-world environments; both types of evaluations are equally important and should be meticulously carried out before full-scale adoption.

Conclusions

We identify the key challenges that researchers face in developing accurate, robust, and usable machine learning models that can create value in clinical radiology practice. These challenges and the techniques to overcome them have been discussed previously in a piecemeal manner in prior research literature. In this study, we re-examined them in the context of medical imaging. By compiling them in the form of a laundry list, we hope to make this research more readily accessible.

Hospital workflows and practices vary widely from one hospital to another, even within the same geography. This increases the difficulty of seamlessly integrating predictive models into hospital workflows. The nonuniformity in workflows also raises the question of whether the reported performance of a model is reproducible in a different clinical context. This is an ongoing research, and satisfactory solutions are yet to be found.

The ultimate objective of diagnostic machine learning models is to improve patient outcomes. However, improvement in diagnostic performance does not, by itself, cause an improvement in patient outcomes [31]. Radiological diagnosis is only one of the many steps that eventually leads to treatment. Therefore, a computerized diagnostic system must be placed appropriately in the workflow. How the system presents the results to the reporting radiologist and what action the radiologist takes on receiving them are important factors that influence the usefulness of the system in practice.

On the one hand, medical imaging is a broad and complex field that encompasses numerous imaging modalities, pathological conditions, and diagnostic protocols. On the other hand, machine learning is an active area of research with thousands of new techniques published every year. The combined diversity of both fields along with nonuniform hospital practices, regulatory restrictions on data sharing, and lack of standardized reporting of results make it difficult to clearly assess the role and potential of machine learning applications in medical imaging. We believe that machine learning has great potential in improving diagnostic accuracy, lowering reporting times, reducing radiologist workloads, and ultimately improving the delivery of health care. To realize this potential, however, a concerted across-the-board effort will be required from physicians, radiologists, patients, hospital administrators, data scientists, software developers, and other stakeholders.

Authors' Contributions

VK is the primary author of the paper and was responsible for conceiving the topic, reviewing the research literature, and writing the manuscript. MG assisted in surveying the research literature and writing some sections of the manuscript. AK is a senior radiologist who validated the manuscript from a clinical radiology perspective and assisted in editing the manuscript.

Conflicts of Interest

None declared.

References

1. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. Br Med J 2017 Oct 11;359:j4683. [doi: 10.1136/bmj.j4683] [Medline: 29021184]

2. Nakajima Y, Yamada K, Imamura K, Kobayashi K. Radiologist supply and workload: international comparison--Working Group of Japanese College of Radiology. Radiat Med 2008 Oct;26(8):455-465. [doi: 10.1007/s11604-008-0259-2] [Medline: 18975046]
3. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
4. Maretíc Z. Cnidarismus nudorum: A new epidemiological and clinical entity. Dermatologica 1986;172(2):123-125. [Medline: 2868931]
5. Shen D, Wu G, Suk H. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017 Jun 21;19:221-248 [FREE Full text] [doi: 10.1146/annurev-bioeng-071516-044442] [Medline: 28301734]
6. Foody G, McCulloch MB, Yates WB. The effect of training set size and composition on artificial neural network classification. Int J Remote Sens 2007 May 03;16(9):1707-1723. [doi: 10.1080/01431169508954507]
7. Kerlikowske K, Grady D, Barclay J, Frankel SD, Ominsky SH, Sickles EA, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst 1998 Dec 02;90(23):1801-1809. [doi: 10.1093/jnci/90.23.1801] [Medline: 9839520]
8. Moifo B, Pefura-Yone EW, Nguefack-Tsague G, Gharingam ML, Tapouh JR, Kengne A, et al. Inter-observer variability in the detection and interpretation of chest x-ray anomalies in adults in an endemic tuberculosis area. O J Med Imaging 2015 Sep;05(03):143-149. [doi: 10.4236/ojmi.2015.53018]
9. He H, Garcia E. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009 Sep;21(9):1263-1284. [doi: 10.1109/TKDE.2008.239]
10. Pooch E, Ballester P, Barros R. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. arXiv.org. 2020. URL: http://arxiv.org/abs/1909.01940 [accessed 2021-08-16]
11. Widmer G, Kubat M. Learning in the presence of concept drift and hidden contexts. Mach Learn 1996 Apr;23(1):69-101. [doi: 10.1007/bf00116900]
12. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. arXiv.org. 2014. URL: https://arxiv.org/abs/1312.6199 [accessed 2021-08-16]
13. Steinhardt J, Koh P, Liang P. Certified defenses for data poisoning attacks. arXiv.org. 2017. URL: https://arxiv.org/abs/1706.03691 [accessed 2021-08-16]
14. Biggio B, Corona I, Maiorca D. Evasion attacks against machine learning at test time. In: Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer; 2013.
15. Brunese L, Mercaldo F, Reginelli A, Santone A. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. Comput Methods Programs Biomed 2020 Nov;196:105608 [FREE Full text] [doi: 10.1016/j.cmpb.2020.105608] [Medline: 32599338]
16. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
17. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: a radiologist's guide. Radiology 2019 Mar;290(3):590-606. [doi: 10.1148/radiol.2018180547] [Medline: 30694159]
18. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers R. ChestX-Ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jul 21-26, 2017; Honolulu, HI, USA. [doi: 10.1109/cvpr.2017.369]
19. Bustos A, Pertusa A, Salinas J, de la Iglesia-Vayá M. PadChest: a large chest x-ray image dataset with multi-label annotated reports. Med Image Anal 2020 Dec;66:101797. [doi: 10.1016/j.media.2020.101797] [Medline: 32877839]
20. Johnson A, Pollard T, Greenbaum N, Lungren M, Deng C, Peng Y, et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv.org. 2019. URL: http://arxiv.org/abs/1901.07042 [accessed 2021-08-16]
21. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019 Jul 6;6(1):60. [doi: 10.1186/s40537-019-0197-0]
22. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. Med Image Anal 2019 Dec;58:101552. [doi: 10.1016/j.media.2019.101552] [Medline: 31521965]
23. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health 2014 Nov 05;14:1144 [FREE Full text] [doi: 10.1186/1471-2458-14-1144] [Medline: 25377061]
24. Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR). Switzerland: Springer; 2017.
25. Annas GJ. HIPAA regulations - a new era of medical-record privacy? N Engl J Med 2003 Apr 10;348(15):1486-1490. [doi: 10.1056/NEJMlim035027] [Medline: 12686707]
26. Chik WB. The Singapore Personal Data Protection Act and an assessment of future trends in data privacy reform. Comput Law Secur Rev 2013 Oct;29(5):554-575. [doi: 10.1016/j.clsr.2013.07.010]
27. McMahan H, Moore E, Ramage D, Hampson S, Arcas B. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. 2017

Presented at: 20th International Conference on Artificial Intelligence and Statistics; Fort Lauderdale, Florida, USA; May 9-11, 2017 URL: http://proceedings.mlr.press/v54/mcmahan17a.html

28. Vepakomma P, Gupta O, Swedish T, Raskar R. Split learning for health: Distributed deep learning without sharing raw patient data. arXiv.org. 2018. URL: http://arxiv.org/abs/1812.00564 [accessed 2021-08-16]
29. Gawali M, Suryavanshi S, CS A, Madaan H, Gaikwad A, Bhanu Prakash KN, et al. Comparison of privacy-preserving distributed deep learning methods in healthcare. In: Proceedings of the Annual Conference on Medical Image Understanding and Analysis. 2021 Presented at: Annual Conference on Medical Image Understanding and Analysis; July 12-14, 2021; Oxford, United Kingdom. [doi: 10.1007/978-3-030-80432-9_34]
30. Kulkarni V, Kulkarni M, Pant A. Survey of personalization techniques for federated learning. In: Proceedings of the Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). 2020 Presented at: Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4); July 27-28, 2020; London, UK. [doi: 10.1109/worlds450073.2020.9210355]
31. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018 Mar;286(3):800-809. [doi: 10.1148/radiol.2017171920] [Medline: 29309734]
32. Zech J, Pain M, Titano J, Badgeley M, Schefflein J, Su A, et al. Natural language-based machine learning models for the annotation of clinical radiology reports. Radiology 2018 May;287(2):570-580. [doi: 10.1148/radiol.2018171093] [Medline: 29381109]
33. Smit A, Jain S, Rajpurkar P, Pareek A, Ng A, Lungren M. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020 Presented at: Conference on Empirical Methods in Natural Language Processing (EMNLP); Nov 16-20, 2020; Punta Cana URL: https://aclanthology.org/2020.emnlp-main.117.pdf [doi: 10.18653/v1/2020.emnlp-main.117]
34. Pons E, Braun LM, Hunink MG, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016 May;279(2):329-343. [doi: 10.1148/radiol.16142770] [Medline: 27089187]
35. Cheplygina V, de Bruijne M, Pluim JP. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal 2019 May;54:280-296. [doi: 10.1016/j.media.2019.03.009] [Medline: 30959445]
36. Feyjie A, Azad R, Pedersoli M, Kauffman C, Ayed I, Dolz J. Semi-supervised few-shot learning for medical image segmentation. arXiv.org. 2020. URL: http://arxiv.org/abs/2003.08462 [accessed 2021-08-16]
37. Cheng Q, Zhang Q, Fu P, Tu C, Li S. A survey and analysis on automatic image annotation. Pattern Recognit 2018 Jul;79:242-259. [doi: 10.1016/j.patcog.2018.02.017]
38. Brady AP. Radiology reporting-from Hemingway to HAL? Insights Imaging 2018 Apr;9(2):237-246 [FREE Full text] [doi: 10.1007/s13244-018-0596-3] [Medline: 29541954]
39. Olatunji T, Yao L, Covington B, Rhodes A, Upton A. Caveats in generating medical imaging labels from radiology reports. arXiv.org. 2019. URL: http://arxiv.org/abs/1905.02283 [accessed 2021-08-16]
40. Rosenkrantz AB, Duszak R, Babb JS, Glover M, Kang SK. Discrepancy rates and clinical impact of imaging secondary interpretations: a systematic review and meta-analysis. J Am Coll Radiol 2018 Sep;15(9):1222-1231. [doi: 10.1016/j.jacr.2018.05.037] [Medline: 30031614]
41. Brouwer CL, Steenbakkers RJ, van den Heuvel E, Duppen JC, Navran A, Bijl HP, et al. 3D variation in delineation of head and neck organs at risk. Radiat Oncol 2012 Mar 13;7:32 [FREE Full text] [doi: 10.1186/1748-717X-7-32] [Medline: 22414264]
42. Njeh CF. Tumor delineation: the weakest link in the search for accuracy in radiotherapy. J Med Phys 2008 Oct;33(4):136-140 [FREE Full text] [doi: 10.4103/0971-6203.44472] [Medline: 19893706]
43. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. AAAI-19 2019 Jul 17;33(1):590-597. [doi: 10.1609/aaai.v33i01.3301590]
44. Sakurada S, Hang NT, Ishizuka N, Toyota E, Hung LD, Chuc PT, et al. Inter-rater agreement in the assessment of abnormal chest X-ray findings for tuberculosis between two Asian countries. BMC Infect Dis 2012 Feb 01;12:31 [FREE Full text] [doi: 10.1186/1471-2334-12-31] [Medline: 22296612]
45. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 2011 Feb;38(2):915-931 [FREE Full text] [doi: 10.1118/1.3528204] [Medline: 21452728]
46. Majkowska A, Mittal S, Steiner DF, Reicher JJ, McKinney SM, Duggan GE, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology 2020 Feb;294(2):421-431. [doi: 10.1148/radiol.2019191293] [Medline: 31793848]
47. Raykar V, Yu S, Zhao L, Jerebko A, Florin C, Valadez G, et al. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: Proceedings of the 26th Annual International Conference on Machine Learning. 2009 Presented at: ICML '09: 26th Annual International Conference on Machine Learning; Jun 14-18, 2009; Montreal Quebec Canada. [doi: 10.1145/1553374.1553488]

48. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019 Mar 19;6(1):27. [doi: 10.1186/s40537-019-0192-5]
49. Liu Y, Yu X, Huang JX, An A. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Inf Process Manag 2011 Jul;47(4):617-631. [doi: 10.1016/j.ipm.2010.11.007]
50. Kim J, Kim J. The impact of imbalanced training data on machine learning for author name disambiguation. Scientometrics 2018 Jul;117(3-4):511-526. [doi: 10.1007/s11192-018-2865-9]
51. Chen H, Xiong F, Wu D, Zheng L, Peng A, Hong X, et al. Assessing impacts of data volume and data set balance in using deep learning approach to human activity recognition. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2017 Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Nov 13-16, 2017; Kansas City, MO, USA. [doi: 10.1109/bibm.2017.8217821]
52. Chawla N. Data mining for imbalanced datasets: an overview. In: Data Mining and Knowledge Discovery Handbook. Boston, MA: Springer; 2005:853-867.
53. Luque A, Carrasco A, Martín A, de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognit 2019 Jul;91:216-231. [doi: 10.1016/j.patcog.2019.02.023]
54. Drummond C, Holte R. C4. 5, class imbalance and cost sensitivity: why under-sampling beats over-sampling. In: Proceedings of the International Conference on Machine Learning (ICML 2003) Workshop on Learning from Imbalanced Data Sets II. 2003 Presented at: International Conference on Machine Learning (ICML 2003) Workshop on Learning from Imbalanced Data Sets II; Jul 21, 2003; Washington, DC, USA URL: https://www.site.uottawa.ca/~nat/Workshop2003/drummondc.pdf
55. Batista G, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor Newsl 2004 Jun 01;6(1):20-29. [doi: 10.1145/1007730.1007735]
56. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]
57. Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Comput Intell 2004 Feb;20(1):18-36. [doi: 10.1111/j.0824-7935.2004.t01-1-00228.x]
58. Yuan X, Xie L, Abouelenien M. A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. Pattern Recognit 2018 May;77:160-172. [doi: 10.1016/j.patcog.2017.12.017]
59. Wei Q, Shi B, Lo J, Carin L, Ren Y, Hou R. Anomaly detection for medical images based on a one-class classification. In: Proceedings of the Conference on Medical Imaging 2018: Computer-Aided Diagnosis. 2018 Presented at: Conference on Medical Imaging 2018: Computer-Aided Diagnosis; Feb 27, 2018; Houston, Texas, United States. [doi: 10.1117/12.2293408]
60. Ruff L, Vandermeulen R, Goernitz N, Deecke L, Siddiqui S, Binder A, et al. Deep one-class classification. In: Proceedings of the 35th International Conference on Machine Learning. 2018 Presented at: 35th International Conference on Machine Learning; Jul 10-15, 2018; Stockholm Sweden URL: http://proceedings.mlr.press/v80/ruff18a.html
61. Elkan C. The foundations of cost-sensitive learning. In: Proceedings of the 17th international joint conference on Artificial intelligence. 2001 Presented at: IJCAI'01: 17th international joint conference on Artificial intelligence; Aug 4, 2001; Seattle WA USA. [doi: 10.5555/1642194.1642224]
62. Sun Y, Kamel MS, Wong AK, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit 2007 Dec;40(12):3358-3378. [doi: 10.1016/j.patcog.2007.04.009]
63. Maloof M. Learning when data sets are imbalanced and when costs are unequal and unknown. In: Proceedings of the ICML'2003 Workshop: Learning from Imbalanced Data Sets II. 2003 Presented at: ICML'2003 Workshop: Learning from Imbalanced Data Sets II; Aug 21, 2003; Washington, DC URL: https://www.site.uottawa.ca/~nat/Workshop2003/maloof-icml03-wids.pdf
64. Joshi M, Kumar V, Agarwal R. Evaluating boosting algorithms to classify rare classes: comparison and improvements. In: Proceedings of the 2001 IEEE International Conference on Data Mining. 2001 Presented at: IEEE International Conference on Data Mining; Nov 29-Dec 2, 2001; San Jose, CA, USA. [doi: 10.1109/icdm.2001.989527]
65. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag 2009 Jul;45(4):427-437. [doi: 10.1016/j.ipm.2009.03.002]
66. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. 2006 Jun Presented at: ICML '06: 23rd international conference on Machine learning; Jun 25-29, 2006; Pittsburgh Pennsylvania USA. [doi: 10.1145/1143844.1143874]
67. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015 Mar 4;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
68. Drummond C, Holte R. What ROC Curves Can’t Do (and Cost Curves Can). URL: https://www.site.uottawa.ca/~nat/Courses/csi5388/Presentations/cost_curves.pdf [accessed 2021-08-16]
69. Drummond C, Holte RC. Cost curves: an improved method for visualizing classifier performance. Mach Learn 2006 May 8;65(1):95-130. [doi: 10.1007/s10994-006-8199-5]
70. Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest x-ray classification. Sci Rep 2019 Apr 23;9(1):6381 [FREE Full text] [doi: 10.1038/s41598-019-42294-8] [Medline: 31011155]

71. Rajpurkar P, Joshi A, Pareek A, Chen P, Kiani A, Irvin J, et al. CheXpedition: investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting. arXiv.org. 2020. URL: http://arxiv.org/abs/2002.11379 [accessed 2021-08-16]
72. 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 Nov 6;15(11):e1002683 [FREE Full text] [doi: 10.1371/journal.pmed.1002683] [Medline: 30399157]
73. Elgendi M, Nasir MU, Tang Q, Smith D, Grenier J, Batte C, et al. The effectiveness of image augmentation in deep learning networks for detecting COVID-19: a geometric transformation perspective. Front Med (Lausanne) 2021 Mar 1;8:629134 [FREE Full text] [doi: 10.3389/fmed.2021.629134] [Medline: 33732718]
74. Glocker B, Robinson R, Castro D, Dou Q, Konukoglu E. Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects. arXiv.org. 2019. URL: http://arxiv.org/abs/1910.04597 [accessed 2021-08-16]
75. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. Mach Learn 2009 Oct 23;79:151-175. [doi: 10.1007/s10994-009-5152-4]
76. Wang M, Deng W. Deep visual domain adaptation: a survey. Neurocomputing 2018 Oct 27;312:135-153. [doi: 10.1016/j.neucom.2018.05.083]
77. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. In: Domain Adaptation in Computer Vision Applications. Basel, Switzerland: Springer; Sep 13, 2017.
78. Long M, Zhu H, Wang J, Jordan M. Unsupervised domain adaptation with residual transfer networks. arXiv.org. 2017. URL: http://arxiv.org/abs/1602.04433 [accessed 2021-08-16]
79. Tzeng E, Hoffman J, Saenko K, Darrell T. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jul 21-26, 2017; Honolulu, HI, USA. [doi: 10.1109/cvpr.2017.316]
80. Dou Q, Castro DD, Kamnitsas K, Glocker B. Domain generalization via model-agnostic learning of semantic features. arXiv.org. 2019. URL: https://arxiv.org/abs/1910.13580 [accessed 2021-08-16]
81. Bousmalis K, Trigeorgis G, Silberman N, Krishnan D, Erhan D. Domain separation networks. arXiv.org. 2016. URL: http://arxiv.org/abs/1608.06019 [accessed 2021-08-16]
82. Li H, Pan S, Wang S, Kot A. Domain generalization with adversarial feature learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Salt Lake City, UT, USA; Jun 18-23, 2018. [doi: 10.1109/cvpr.2018.00566]
83. Muandet K, Balduzzi D, Schölkopf B. Domain generalization via invariant feature representation. In: Proceedings of the 30th International Conference on Machine Learning. 2013 Presented at: 30th International Conference on Machine Learning; Jun 16-21, 2013; Atlanta, Georgia URL: http://proceedings.mlr.press/v28/muandet13.html
84. Volpi R, Namkoong H, Sener O, Duchi J, Murino V, Savarese S. Generalizing to unseen domains via adversarial data augmentation. In: Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). 2018 Presented at: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018); Dec 2-8, 2018; Montréal, Canada URL: https://papers.nips.cc/paper/2018/file/1d94108e907bb8311d8802b48fd54b4a-Paper.pdf
85. Peng X, Huang Z, Sun X, Saenko K. Domain agnostic learning with disentangled representations. arXiv.org. 2019. URL: http://arxiv.org/abs/1904.12347 [accessed 2021-08-16]
86. Žliobaitė I. Learning under concept drift: an overview. arXiv.org. 2010. URL: http://arxiv.org/abs/1010.4784 [accessed 2021-08-16]
87. Wang S, Minku LL, Yao X. A systematic study of online class imbalance learning with concept drift. IEEE Trans Neural Netw Learn Syst 2018 Oct;29(10):4802-4821. [doi: 10.1109/tnnls.2017.2771290]
88. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv 2014 Apr;46(4):1-37. [doi: 10.1145/2523813]
89. Žliobaitė I, Pechenizkiy M, Gama J. An overview of concept drift applications. In: Big Data Analysis: New Algorithms for a New Society. New York City: Springer International; 2016.
90. Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv.org. URL: http://arxiv.org/abs/1412.6572 [accessed 2021-08-16]
91. Moosavi-Dezfooli S, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016; Las Vegas, NV, USA. [doi: 10.1109/cvpr.2016.282]
92. Qayyum A, Qadir J, Bilal M, Al-Fuqaha A. Secure and robust machine learning for healthcare: a survey. IEEE Rev Biomed Eng 2020 Jul 31;14:156-180 [FREE Full text] [doi: 10.1109/rbme.2020.3013489]
93. Huang R, Xu B, Schuurmans D, Szepesvari C. Learning with a strong adversary. arXiv.org. 2016. URL: http://arxiv.org/abs/1511.03034 [accessed 2021-08-16]
94. Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples. arXiv.org. 2015. URL: http://arxiv.org/abs/1412.5068 [accessed 2021-08-16]

95. Xu W, Evans D, Qi Y. Feature squeezing: detecting adversarial examples in deep neural networks. In: Proceedings of the Network and Distributed Systems Security Symposium (NDSS). 2018 Presented at: Network and Distributed Systems Security Symposium (NDSS); Feb 18-21, 2018; San Diego, CA. [doi: 10.14722/ndss.2018.23198]
96. Gao J, Wang B, Lin Z, Xu W, Qi Y. DeepCloak: masking deep neural network models for robustness against adversarial samples. arXiv.org. 2017. URL: http://arxiv.org/abs/1702.06763 [accessed 2021-08-16]
97. Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. In: Proceedings of the IEEE Symposium on Security and Privacy (SP). 2016 Presented at: 2016 IEEE Symposium on Security and Privacy (SP); May 22-26, 2016; San Jose, CA, USA. [doi: 10.1109/sp.2016.41]
98. Katz G, Barrett C, Dill D, Julian K, Kochenderfer M. Reluplex: an efficient SMT solver for verifying deep neural networks. In: Computer Aided Verification. Basel, Switzerland: Springer; 2017.
99. Ross A, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. arXiv.org. 2017. URL: http://arxiv.org/abs/1711.09404 [accessed 2021-08-16]
100. Bradshaw J, Matthews A, Ghahramani Z. Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks. arXiv.org. 2017. URL: http://arxiv.org/abs/1707.02476 [accessed 2021-08-16]
101. Nguyen L, Wang S, Sinha A. A learning and masking approach to secure learning. In: Decision and Game Theory for Security. Basel, Switzerland: Springer International; 2018.
102. Metzen J, Genewein T, Fischer V, Bischoff B. On detecting adversarial perturbations. arXiv.org. 2017. URL: http://arxiv.org/abs/1702.04267 [accessed 2021-08-16]
103. Lu J, Issaranon T, Forsyth D. SafetyNet: detecting and rejecting adversarial examples robustly. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017 Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); Oct 22-29, 2017; Venice, Italy. [doi: 10.1109/iccv.2017.56]
104. Gopinath D, Katz G, Pasareanu C, Barrett C. DeepSafe: a data-driven approach for checking adversarial robustness in neural networks. arXiv.org. 2020. URL: http://arxiv.org/abs/1710.00486 [accessed 2021-08-16]
105. Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P. Ensemble adversarial training: attacks and defenses. arXiv.org. 2020. URL: http://arxiv.org/abs/1705.07204 [accessed 2021-08-16]
106. Song Y, Kim T, Nowozin S, Ermon S, Kushman N. PixelDefend: leveraging generative models to understand and defend against adversarial examples. arXiv.org. 2018. URL: http://arxiv.org/abs/1710.10766 [accessed 2021-08-16]
107. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019 Mar 22;363(6433):1287-1289 [FREE Full text] [doi: 10.1126/science.aaw4399] [Medline: 30898923]
108. Chakraborty A, Alam M, Dey V, Chattopadhyay A, Mukhopadhyay D. Adversarial attacks and defences: a survey. arXiv.org. 2018. URL: http://arxiv.org/abs/1810.00069 [accessed 2021-08-16]
109. Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access 2018 Feb 19;6:14410-14430. [doi: 10.1109/access.2018.2807385]
110. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017 Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); Oct 22-29, 2017; Venice, Italy. [doi: 10.1109/iccv.2017.74]
111. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Aug Presented at: KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco California USA. [doi: 10.1145/2939672.2939778]
112. Buhrmester V, Münch D, Arens M. Analysis of explainers of black box deep neural networks for computer vision: a survey. arXiv.org. 2019. URL: http://arxiv.org/abs/1911.12116 [accessed 2021-08-16]
113. Campolo A, Sanfilippo M, Whittaker M, Crawford K. AI now 2017 report. AI Now. 2017. URL: https://ainowinstitute.org/AI_Now_2017_Report.pdf [accessed 2021-08-16]
114. Milli S, Schmidt L, Dragan A, Hardt M. Model reconstruction from model explanations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019 Presented at: FAT* '19: Conference on Fairness, Accountability, and Transparency; Jan 29-31, 2019; Atlanta GA USA. [doi: 10.1145/3287560.3287562]
115. Shokri R, Strobel M, Zick Y. On the privacy risks of model explanations. arXiv.org. 2021. URL: http://arxiv.org/abs/1907.00164 [accessed 2021-08-16]
116. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A 2020 Jun 09;117(23):12592-12594 [FREE Full text] [doi: 10.1073/pnas.1919012117] [Medline: 32457147]
117. Verma S, Rubin J. Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. 2018 Presented at: FairWare '18: International Workshop on Software Fairness; May 29, 2018; Gothenburg Sweden. [doi: 10.1145/3194770.3194776]
118. Corbett-Davies S, Goel S. The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv.org. 2018. URL: http://arxiv.org/abs/1808.00023 [accessed 2021-08-16]

119. Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger K. On fairness and calibration. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Dec Presented at: NIPS'17: 31st International Conference on Neural Information Processing Systems; December 4 - 9, 2017; Long Beach California USA p. 5684-5693. [doi: 10.5555/3295222.3295319]
120. Benjamin R. Assessing risk, automating racism. Science 2019 Oct 25;366(6464):421-422. [doi: 10.1126/science.aaz3873] [Medline: 31649182]
121. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 2019 Mar;20(3):405-410 [FREE Full text] [doi: 10.3348/kjr.2019.0025] [Medline: 30799571]
122. Mahajan V, Venugopal VK, Murugavel M, Mahajan H. The algorithmic audit: working with vendors to validate radiology-AI algorithms-how we do it. Acad Radiol 2020 Jan;27(1):132-135. [doi: 10.1016/j.acra.2019.09.009] [Medline: 31818381]
123. Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br Med J 2015 Jan 07;350:g7594. [doi: 10.1136/bmj.g7594] [Medline: 25569120]

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
NLP: natural language processing

Edited by C Lovis; submitted 15.03.21; peer-reviewed by V Gupta, G Liu; comments to author 17.06.21; revised version received 29.06.21; accepted 10.07.21; published 09.09.21

Please cite as:
Kulkarni V, Gawali M, Kharat A
Key Technology Considerations in Developing and Deploying Machine Learning Models in Clinical Radiology Practice
JMIR Med Inform 2021;9(9):e28776
URL: https://medinform.jmir.org/2021/9/e28776
doi: 10.2196/28776
PMID: