European Radiology
https://doi.org/10.1007/s00330-019-06564-3

CHEST

Deep learning: definition and perspectives for thoracic imaging

Guillaume Chassagnon[1,2] · Maria Vakalopolou[2] · Nikos Paragios[2,3] · Marie-Pierre Revel[1]

Received: 10 July 2019 / Revised: 22 October 2019 / Accepted: 30 October 2019

Abstract

Relevance and penetration of machine learning in clinical practice is a recent phenomenon with multiple applications being currently under development. Deep learning—and especially convolutional neural networks (CNNs)—is a subset of machine learning, which has recently entered the field of thoracic imaging. The structure of neural networks, organized in multiple layers, allows them to address complex tasks. For several clinical situations, CNNs have demonstrated superior performance as compared with classical machine learning algorithms and in some cases achieved comparable or better performance than clinical experts. Chest radiography, a high-volume procedure, is a natural application domain because of the large amount of stored images and reports facilitating the training of deep learning algorithms. Several algorithms for automated reporting have been developed. The training of deep learning algorithm CT images is more complex due to the dimension, variability, and complexity of the 3D signal. The role of these methods is likely to increase in clinical practice as a complement of the radiologist's expertise. The objective of this review is to provide definitions for understanding the methods and their potential applications for thoracic imaging.

Key Points

- *Deep learning outperforms other machine learning techniques for number of tasks in radiology.*
- *Convolutional neural network is the most popular deep learning architecture in medical imaging.*
- *Numerous deep learning algorithms are being currently developed; some of them may become part of clinical routine in the near future.*

Keywords Machine learning · Deep learning · Lung · Thorax

Abbreviations

AI	Artificial intelligence
CAD	Computer-aided diagnosis
CNNs	Convolutional neural networks
COPD	Chronic obstructive pulmonary disease
CT	Computed tomography
EGFR	Epithelial growth factor receptor
GAN	Generative adversarial neural networks
GPU	Graphic processing unit
NIH	National Institute of Health
PACS	Picture archiving and communication systems
RNN	Recurrent neural networks
SVM	Support vector machine

✉ Marie-Pierre Revel
marie-pierre.revel@aphp.fr

[1] Service de Radiologie A, Radiology Department, Groupe Hospitalier Cochin Broca Hôtel-Dieu, AP-HP, Université Paris Descartes, 27 Rue du Faubourg Saint-Jacques, 75014 Paris, France

[2] Center for Visual Computing, Ecole CentraleSupelec, 3 Rue Joliot Curie, 91190 Gif-sur-Yvette, France

[3] Pépinière Paris Santé Cochin, TheraPanacea, 27 Rue Faubourg Saint Jacques, 75014 Paris, France

Introduction

The term "machine learning" was introduced in 1959 by Arthur L. Samuel, who designed the first program for the game of checkers [1]. Machine learning is a subset of methods of artificial intelligence (AI). Its aim is to develop algorithms that learn interpretation principles from training samples, and apply them to new data from the same domain to make informed decisions. Deep learning—a subset of machine learning—has recently become a hot topic in radiology.

Published online: 06 December 2019

Indeed, deep learning for a specific class of problems has been shown to outperform other machine learning methods, allowing the creation of models that perform as well or even better than humans [2, 3]. Such a revolution was driven from the increasing availability of large datasets, and high computing capacity of graphic processing units (GPU), even though there were also algorithmic and mathematical progresses in neural network learning.

Machine learning is especially relevant for image interpretation. It adopts an evidence-driven concept where the underlying decision process is very different from one traditionally adopted by radiologists. The objective of this review is to help readers to become familiar with the methods and their potentials, and to report current and future developments applied to thoracic imaging. This article will mainly focus on models commonly used in radiology and especially on a specific type of deep learning networks, the convolutional neural network (CNN).

Terminology

In order to understand machine learning, the first important step is to understand the different terms being used.

In 1959, Arthur Samuel defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed" [1].

Conventional programing relies on a logic that is introduced during its conception and does not change. Machine learning applies a different principle where the behavior of the program changes according to the training data. It can generate systems that are able to automatically learn from the available data, without "being explicitly programmed." Both classical machine learning methods and deep learning methods use optimization algorithms during training. The main difference is that classical machine learning methods require the selection of image features beforehand, while the deep learning ones generate these features during training.

Among various algorithms usually used in machine learning, neural networks are designed to mimic the way the human brain processes information. In brief, successions of simple operations—mimicking the way neurons behave—are used to treat the information. Each neuron (formal neuron) processes part of the signal. The composition of these processes is used to build the decision algorithm, also called the model.

Deep learning refers to deep neural network, which is a specific configuration where neurons are organized in multiple successive layers [4]. The increase of layers improves the expression power and performance of these methods and could produce higher level of abstraction [3]. Deep learning currently represents the most advanced machine learning technique for a variety of high-level tasks and applications, especially for problems involving large structured training datasets, which is the case for chest radiograph interpretation. In the context of radiology, its goal is to develop algorithms and tools for the automated processing, analysis, and understanding of digital images towards reproducing the human visual perception system.

The term CAD (computer-aided diagnosis) is a generic term encompassing various mathematical methods not limited to deep learning [5]. For thoracic imaging, the most prominent application refers to lung nodule diagnosis. This includes CAD for detection, named CADe, and CAD for characterization, named CADx, used to evaluate the probability of malignancy. Some CADs combine both tasks [6].

Radiomics is another popular research direction, relying on more traditional machine learning tools, with some recent development exploiting deep learning methods. The objective is to determine imaging features of various complexities, which are invisible to the human eye, in order to establish correlations with clinical outcomes. Classical machine learning algorithms are usually using three different categories of features: morphological features such as shape, volume, and diameter; image features or first-order features such as histogram, kurtosis, and mean values; and textural features (higher order features) including co-occurrence of patterns and filter responses. These features are extracted and analyzed to be used for classification purposes (is the nodule benign or malignant?), for quantification (what is the degree of severity of this bronchial disease?) [7], or for prognosis, response to treatment, or correlation with other clinical or biological biomarkers. There are many applications of radiomics in thoracic oncology, such as discriminating adenocarcinoma from squamous cell carcinoma [8], predicting lung adenocarcinoma invasiveness [9] or epithelial growth factor receptor (EGFR) mutation [10], linking the tumor "radiomics phenotype" and the tumor genotype [11], or predicting response to treatment [12]. However, a recent study evaluating 77 articles reported a mean radiomics quality score, a metric evaluating the validity and completeness of radiomics studies, of only 26.1% and concluded that the overall scientific quality of radiomics studies was insufficient [13].

Main concepts regarding machine learning algorithms

Types of algorithms

Machine learning algorithms can be categorized into three main groups: supervised, semi-supervised, or unsupervised algorithms. Algorithms based on supervision rely on samples with annotations provided by clinical experts, which will be used for training. Supervised learning algorithms can be trained for classification tasks, such as to the presence or

absence of disease or anomaly, or for regression tasks, for instance to provide a severity score or a prognosis.

Recently, semi-supervised methods have emerged which combine annotated and non-annotated data. In this setting, algorithms learn progressively through a better exploitation of the non-annotated data. Reinforcement learning is a typical example of semi-supervised learning [14].

Conversely, algorithms based on unsupervised learning do not involve human intervention. Clustering is the most representative example, where the objective is to group samples into homogeneous subpopulations, like, for example, to identify different chronic obstructive pulmonary disease (COPD) phenotypes. The performance of unsupervised algorithms is often lower than the one achieved with supervised techniques.

Data annotation

Supervised and semi-supervised algorithms rely on annotated data. There are different types of annotations depending on the task that the algorithm seeks to address (Fig. 1). For classification tasks (presence or absence of anomaly or disease), images are simply labeled with two (disease positive or negative) or more labels. For instance, chest X-ray 14 database [15] contains around 112,000 chest radiographies labeled as containing one or more of 14 common anomalies such as atelectasis, cardiomegaly, effusion, infiltrates, mass, nodule, pneumonia, and pneumothorax.

Since annotations of large datasets are generated by automated extractions from the reports, it is important to test the accuracy of the automated labeling, which is usually done by comparison with radiologists' annotation on a subset of radiographs [16].

The exact localization of the anomaly is not provided within the image. This is also referred as a weakly annotated database. Even though the annotation does not include exact localization, some algorithm might automatically learn to predict the anatomical position of the anomaly.

The next level of annotation usually refers to a sparse way of providing information [17] or with boundary boxes indicating the regions of interest [18]. Segmentation tasks require the highest level of annotation which consists in contouring/delineating the anomalies on each image. This type of annotation allows building more precise algorithms but is tedious and time-consuming. Such datasets are generally smaller and more difficult to generate.

Database/dataset

For machine learning methods including deep learning, the quality of data is essential and could be even more important than the learning algorithm itself. It guarantees the capacity for the model to perform equally well on cases not seen during training. For obtaining a generalizable model, it is important to have a dataset that is representative of the disease and also representative of the different acquisition techniques. In radiology, datasets must include the different acquisition protocols and the various forms of the evaluated disease and also include examinations from disease-free subjects. A model for lung fibrosis detection should be trained using a dataset reflecting the heterogeneity of lung fibrosis patterns but also including normal CT scans, and CT images acquired on various CT equipment. If the training dataset only contains a unique fibrosis pattern or acquisitions all performed on the same CT unit with the same reconstruction protocol, the risk for the model to be poorly generalizable is high.

The dataset is usually split in three subcategories: training, validation, and testing. The training set—usually corresponding to 60% of the database—is used to train variant versions of the model with different initialization conditions and hyperparameters (these will be defined in upcoming section). Once the models have been trained, their performance is evaluated using 20% of the remaining data, composing what is called the validation dataset. The model with the best performance on the validation dataset is selected. This model is

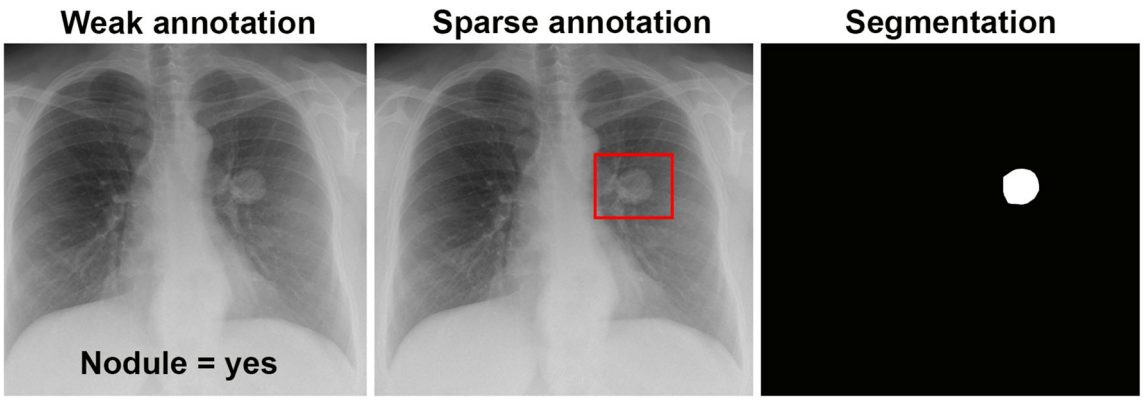

Fig. 1 Different types of annotations. In weak annotation, images are simply labeled (nodule = yes or no) and exact localization of the anomaly is not provided. In sparse annotation, a bounding box is drawn around the nodule, whereas in segmentation, the nodule contour is delineated (white area)

finally evaluated using the last 20% of samples, which were never previously used and compose the test dataset.

An alternative to the dataset splitting between training and validation is the method of k-fold cross-validation which allows training and validating using the entire dataset. This approach is especially useful when the number of cases is limited. It consists of splitting the training and validation sets to several random splits and then using different combinations for training and validation. The average performance of the model on these splits is taken into account to judge the model performance and acceptability.

Data preprocessing

Data preprocessing is not limited to the decomposition of the dataset into training, validation, and test. Although in real life the radiologists need to analyze images acquired with different techniques (machines, dose, slice thickness, pixel size, reconstruction algorithms) and nevertheless make useful comparisons, it is desirable to "normalize" images before feeding them into the deep learning algorithm. Various degrees of preprocessing can be applied, such as normalization of the physical resolution involving slice thickness and voxel size and/or normalization of the gray-level distributions to follow predefined distribution and/or image denoising [19]. Preprocessing is absolutely essential for radiomics studies involving feature selection, which is either selected by humans with traditional machine learning techniques or automatically identified when using deep learning. On the former case, the preprocessing step also involves the choice of image features, among all 3 categories previously described, to feed the algorithm Among these features, the algorithm will select the most prominent ones with respect to the task, using different possible techniques such as random forest, Lasso, SVM, and logistic regression.

Deep learning architectures

In radiology, three architectures are predominantly used.

1. Convolutional neural networks (CNNs) are the most popular ones because they are robust and easy to train [20–22]. They rely on the succession of simple convolution/deconvolution operators at different scales (Fig. 2). Convolution consists in aggregating information from voxels grouped together, through the application of different filters (Fig. 3). The filters differ from one layer to the next and their application generates the input to the subsequent layer. Two main types of convolutional neural networks are very popular for thoracic applications: (i) the fully connected CNNs which are a mixture of convolutional and fully connected layers mainly used for classification purposes and (ii) the fully convolutional CNNs which are composed from only convolutional layers. These architectures are mainly used for segmentation purposes.
2. Recurrent neural networks (RNNs) are used to jointly solve different, interdependent problems, such as detection and characterization of nodules. The network is organized in closed loops rather than in a sequence of operations like CNNs [23]. These loops allow solving the interdependency of tasks. Another application of this architecture is the ability to encode temporal information and deal with dynamic data, for instance enhancement after contrast administration [24].
3. The last class refers to generative adversarial networks (GANs) [25], where during training of the algorithm, information coming from images is combined with a statistical predictor, jointly determining the outcomes. For instance, lung nodules are generally spheroid in shape, and the statistical component of the GAN for a lung nodule detection algorithm will reinforce this condition for the final prediction. Such methods are used when plausibility of the deep learning result is important to consider. Other architectures do not explicitly provide a statistical interpretation of the results. Compared with CNNs and RNNs, GAN architectures are the hardest to train.

Hyperparameters, loss function, and optimization strategy

The term hyperparameter refers to all parameters which are defined before training the algorithm, by opposition to those which will derive from learning. Number and type vary according to the deep learning architecture. Hyperparameters include the number of layers, the learning rate which depends on the loss function and optimization strategy.

The loss function is an important concept to understand. It corresponds to the metrics used by the algorithm during training to test its performance. It quantifies the gap between the prediction by the algorithm and the ground truth given by the expert annotation/label. The objective of any deep learning algorithm is always to minimize its loss function, until the discrepancy between the prediction by the network and the ground truth vanishes. The loss function varies according to the task which is addressed, such as the Dice similarity index loss [26] for segmentation tasks or the log loss for detection and classification tasks.

Several strategies can be used to optimize the loss function during training. The most commonly used is the stochastic gradient descent [27]. Gradient descent methods rely on an iterative process where every iteration allows moving closer to the optimal model. Stochastic gradient descent allows

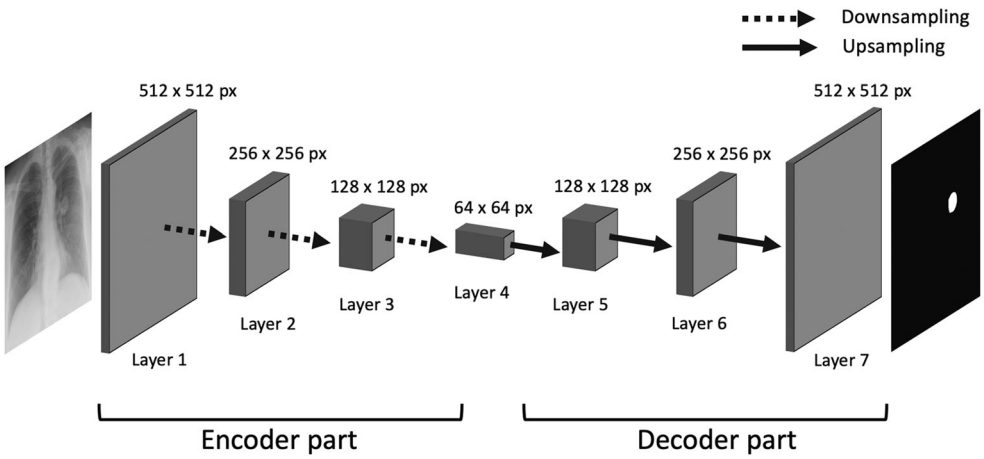

Fig. 2 Illustration of a convolution/deconvolution neural network. The convolution part of the network applies convolution operators at different scales. Scale reduction (downsampling) between each layer is usually obtained by using max-pooling function. Then, the deconvolution part applies deconvolution operators and progressively restores the initial scale of the image (upsampling)

random perturbation of the model that instantly could degrade its performance but finally converge to a better solution.

The learning rate and the number of epochs are also important optimization parameters. The learning rate controls the amount of improvement of the network between two iterations. Low learning rates guaranty improvements, but of marginal importance. High learning rates refer to more unstable models where improvement from one iteration to the next could be more significant, associated with the risk of degrading the overall performance. Epoch is a different concept, which refers to the number of times where the entire training set has been revisited to update the model parameters. A highest number of epochs guaranties better performance on the training set, at the cost of increasing computation complexity as well as the risk of overfitting, by selecting features which are only specific to the training dataset and are poorly generalizable.

Overfitting and underfitting

Overfitting is the situation where the trained model performs very well on the training dataset but fails on the testing set. Overfitting occurs when the model performance keeps improving in the training cohort but decreases in the validation cohort. In other words, the model generates accurate predictions on the training set, but fails to reproduce them on new unseen cases. This can be observed when the training set is not well balanced or when the number of samples is not sufficient. In this situation, it may happen that the algorithm finds an association of features and considers it as relevant for the outcome, while it is only the result of fortuitous feature combinations learned from a non-representative dataset. This association would disappear when using a larger or different sample. Overfitting problems are common with deep learning algorithms containing many layers generating lots of variables to learn (from several hundred to several millions) from small training sets.

Another problem that can be seen during training is underfitting. It occurs when the model fails its adaptation to both training and validation sets. The reasons of underfitting can be multiple. In presence of multiple subpopulations within the training set, models with an insufficient number of parameters will fail to encompass the entire population. Another possible explanation for underfitting relates to the nature of the model that has been chosen for the prediction. For example, when the notion of time is critical for diagnosis (delayed enhancement), a model that ignores this information will most inevitably fail even if the number of parameters is sufficient.

In summary, overfitting is characterized by a high performance on the training dataset contrasting with a poor

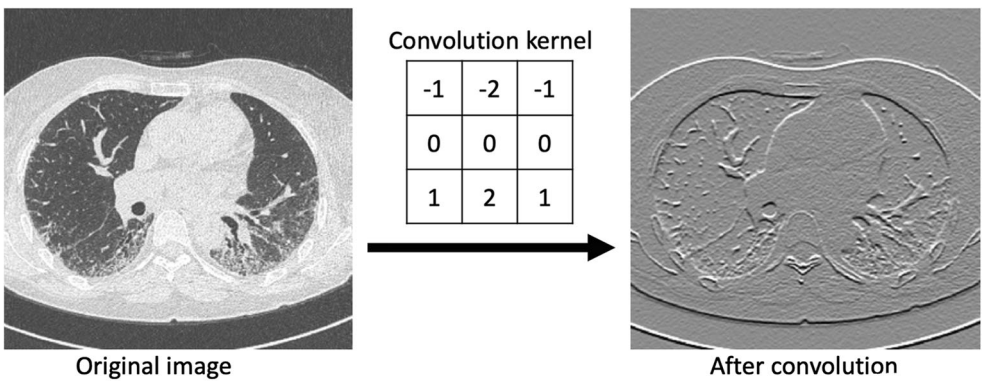

Fig. 3 Example of convolution with Sobel filter to highlight edges on the horizontal direction

performance on the validation dataset, whereas underfitting is characterized by poor performance in both training and validation datasets (Fig. 4).

Perspectives for thoracic imaging

The World Health Organization suggests that two-thirds of the global population lack access to radiology diagnostics [28]. There is a shortage of experts who can interpret X-rays, even when imaging equipment is available, which opens avenues for machine learning applied to thoracic imaging [29].

Chest radiography is an excellent candidate for developing computer-aided automatic interpretation solutions. It consists of 2D images and several billions have already been stored on hospital's picture archiving and communication systems (PACS) and linked to radiological reports. As previously underlined, a large amount of data is essential for supervised learning but images have to be extracted and annotated. Several large databases of annotated chest radiographies are already publicly available for developing research projects. Among them, we already mentioned chest X-ray 8 [15] from NIH with 8 possible labels, secondarily extended to 14 different labels. CheXpert is another large dataset containing 224,316 chest radiographs used for competition for chest X-ray interpretation [16, 30]

Among the numerous studies performed on automated chest X-ray reading, some studies on dedicated tasks have demonstrated the superiority of CNNs over classical machine learning techniques. For the detection of tuberculosis, the deep learning–based algorithm developed by Lakhani et al [31] reached an AUC of 0.99 compared with 0.90 at best for prior studies using classic machine learning methods [32, 33]. Hwang et al developed a deep learning–based algorithm able to distinguish normal and abnormal chest radiograph results, including malignant neoplasm, active tuberculosis, pneumonia, and pneumothorax. The algorithm was trained on a dataset of 54,221 normal chest radiographs and 35,613 with abnormal findings. The algorithm demonstrated significantly higher performance than non-radiology physicians, board-certified radiologists, and thoracic radiologists. These three categories of human readers improved when using the algorithm as second reader [2]. These impressive results at first glance should be tempered by the fact that the X-rays used to compare the model and human performance contained only 1 of the 4 target diseases, which corresponds to the concept of "narrow AI." The performance of the algorithm for detecting the same anomalies when associated with others was not evaluated and might be lower than that of human readers.

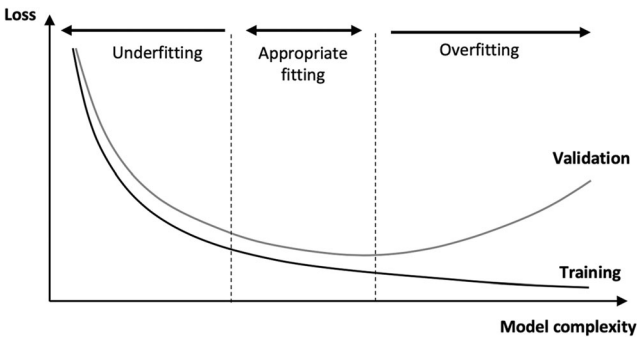

Fig. 4 Underfitting and overfitting. Underfitting is characterized by a high loss in both training and validation datasets, whereas overfitting is characterized by a low loss in the training dataset contrasting with a high loss on the validation dataset

The use of CNNs for CT images is more complex due to the 3D nature and high number of images, and the smaller size of annotated datasets. Despite these difficulties, results are promising and here again CNNs prove superior to classical machine learning methods. For the 2017 Kaggle Data Science Bowl, whose objective was to predict the cancer risk at 1 year, based on lung cancer screening CT examinations, frontrunner teams all used deep learning. The use of deep learning is not restricted to nodule evaluation but can also apply to the detection of emphysema [34] or the detection and quantification of infiltrative lung diseases on CT (Fig. 5) [35].

Deep learning can also be used to predict non-small cell lung cancer genotype from CT images, especially EGFR mutation. Wang et al [36] trained a deep learning algorithm for predicting EGFR mutation on a cohort of 603 patients. They used a 24-layer CNN and pre-trained the first 20 layers on natural images from the ImageNet dataset, before training the 4 remaining layers using 14,926 CT images from wild-type and EGFR-mutated lung adenocarcinomas of the training cohort. The developed algorithm was validated on an external cohort of 241 patients and reached and AUC of 0.81. Deep learning using RNN can also be used to predict lung cancer response to treatment from serial medical imaging [37]. The model was predictive of survival and cancer-specific outcomes, with an AUC of 0.74 for 2-year overall survival.

Conclusion

The major advances of machine learning and in particular deep learning might change the landscape of radiology. Numerous algorithms are being currently developed; part of them might be used for the clinical routine in the coming years. They may assist radiologists for a number of tasks such as detection or characterization of radiological abnormalities, or for prognostic purposes. For detection tasks, human validation by visual confirmation is possible, which is not the case for characterization tasks or prognostic prediction. Who will

Fig. 5 Interstitial lung disease segmentation in a patient with idiopathic pulmonary fibrosis using a deep learning–based tool developed at our research laboratory. **a** Unenhanced CT scan axial transverse image through the lung bases demonstrates ground glass, reticulations, and bronchiolectasis with subpleural predominance. **b** ILD segmentation (red). **c** Combination of ILD (red) and lung (blue) segmentations allowing calculating the volume of diseased lung on CT

take the responsibility for the wrong predictions remains an open question.

Before the introduction of such tools in clinical practice, it is important to understand the associated terms and concepts, the strengths, and the limitations in order to all together become "augmented" radiologists, not overwhelmed radiologists.

Funding Information The authors state that this work has not received any funding.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Pr. MP Revel.

Conflict of interest Pr. N Paragios is an employee of TheraPanacea (Paris, France).

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was not required for this study because this is a review.

Ethical approval Institutional Review Board approval was not required because this is a review.

Methodology
• Review

Glossary

Algorithm	a mathematical model that is applied to data (input) and provide an output in order to solve a specific problem.
Artificial intelligence	a scientific domain that creates algorithms allowing machines to mimic human cognition or human performance on a dataset. Currently, AI algorithms address specific tasks such as tuberculosis diagnosis or pneumonia detection, which is defined as narrow or weak AI. They do not allow the detection of all potential abnormalities (e.g., general AI) as a human reader would do.
CAD (computer-aided diagnosis)	a domain that exploits algorithms derived from artificial intelligence to provide indicators and assist clinical experts in diagnosis.
CADe	CAD developed for a detection task.
CADx	CAD developed for a characterization task.
Classification task	assignment to an input signal (an image) a label from a predefined set of categories (disease or no disease), by mean of a machine learning algorithm.
Cross validation	a statistical method used to estimate the performance of the machine learning algorithm by exploiting various partitions of the data between training and testing.
Convolution	mathematical operator that creates a new value from an input signal (for instance a group of voxels) after modification by another value which acts as a filter. For example, averaging mean density values within a patch of voxels.
Convolutional neural network (CNN)	deep neural network which is based on a sequence of convolutional operations.
Deep learning (= deep neural network)	part of the broader family of machine learning, individualized by specific configuration of neural network organized in multiples

	layers, emulating the human learning approach and increasing the ability to address complex problems. Deep learning networks are iterative methods that propagate information, training their features automatically through gradient-based optimization methods and backpropagation.
Epoch	indicates the number of times the entire dataset has been used during the iterative optimization of the network.
Features	image characteristics which are invisible to the human eye. Three categories of features are used by classical machine learning algorithms: morphological features such as shape, volume, and diameter; first-order features such as histogram, kurtosis, and mean values; and textural features including co-occurrence of patterns and filter responses.
Formal neuron (= artificial neuron)	mathematical function mimicking the architecture of biological neurons.
Fully connected CNNs	variation of CNNs which consists of connecting all the elements of one layer with all the elements of the next one. Fully connected CNNs are used for classification problems (does this chest radiograph contain signs of tuberculosis?).
Fully convolutional CNNs	variation of CNNs which are composed from only convolutional layers. Fully convolutional CCNs are used for segmentation tasks (is this pixel located in a fibrotic area?).
Generalization capability	capacity for a model to maintain its performance when applied to new cases, unseen during training.
Generative adversarial neural network (GAN)	a neural network that combines two subnetworks, one generating hypotheses and another evaluating their likelihood.
Ground truth	refers to the label assigned by the expert or another reference method such as pathology.
Hyperparameters	parameters which control the training process of the algorithm and are defined before training, such as the number of layers and learning rate, among others.
Labeling/annotation	process of allocating ground truth by associating a label to an image.
Loss function	when training and optimizing the algorithm, it quantifies the gap between predictions and ground truth.
Machine learning	a scientific field that gives computers the ability to automatically learn without being explicitly programmed, by relying on sample data, known as "training data," used to make predictions.
Neural network	machine learning algorithm made of a succession of formal neurons.
Overfitting	characterizes algorithms that perform well on the data on which they have been trained but fail to perform equally well on unseen data.
Radiomics	a field of medical imaging that aims to extract features from medical images, for tasks such as characterization or prediction (prognosis, response to treatment, genotype).
Recurrent neural networks (RNN)	a class of neural networks that integrate interdependencies between different tasks using the same data (detection and characterization) or between different data (temporal post-contrast enhancement).
Regression task	process of associating input data with a continuous outcome (for instance survival).
Semi-supervised learning	class of machine learning techniques that learns from annotated data in order to generate their model and improves its performance using the non-annotated ones
Supervised learning	class of machine learning techniques requiring labeled training data in order to generate their model.
Semantic segmentation	process of associating every voxel with a specific label/class, for

	instance diseased or healthy area, which usually requires manual contouring.
Stochastic gradient descent	an iterative method to optimize machine learning methods, very commonly used for deep learning networks.
Test dataset	dataset which is used to evaluate the performance of the final model.
Training dataset	dataset which is used to train the model.
Transfer learning	concept of exporting parameters, principles, and strategies learned from a dataset to another algorithm, which will be trained on another dataset (for example, learning on nonmedical images before applying to chest imaging).
Unsupervised learning	the class of machine learning techniques that seeks to determine patterns or clusters with similar properties (= phenotypes for instance) from unlabeled data. It usually uses techniques different from deep learning.
Underfitting	inability of an algorithm to perform well on both training and test datasets.
Validation dataset	dataset which is used to determine among different variants of the trained model, the optimal model that should be selected for testing on the remaining unseen cases (test dataset).

References

1. Samuel AL (1959) Some studies in machine learning using the game of checkers. IBM J Res Dev 3:210–229. https://doi.org/10.1147/rd.33.0210
2. Hwang EJ, Park S, Jin K-N et al (2019) Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. JAMA Netw Open 2:e191095. https://doi.org/10.1001/jamanetworkopen.2019.1095
3. Ardila D, Kiraly AP, Bharadwaj S et al (2019) End-to-end lung cancer screening with three-dimensional deep learning on lowdose chest computed tomography. Nat Med 25:954–961. https://doi.org/10.1038/s41591-019-0447-x
4. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
5. Suzuki K (2012) A review of computer-aided diagnosis in thoracic and colonic imaging. Quant Imaging Med Surg 2:14
6. Firmino M, Angelo G, Morais H, Dantas MR, Valentim R (2016) Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. Biomed Eng Online 15:2. https://doi.org/10.1186/s12938-015-0120-7
7. Gillies RJ, Kinahan PE, Hricak H (2015) Radiomics: images are more than pictures, they are data. Radiology 278:563–577. https://doi.org/10.1148/radiol.2015151169
8. Zhu X, Dong D, Chen Z et al (2018) Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer. Eur Radiol 28:2772–2778. https://doi.org/10.1007/s00330-017-5221-1
9. Fan L, FangM LZ et al (2019) Radiomics signature: a biomarker for the preoperative discrimination of lung invasive adenocarcinoma manifesting as a ground-glass nodule. Eur Radiol 29:889–897. https://doi.org/10.1007/s00330-018-5530-z
10. Jia T-Y, Xiong J-F, Li X-Y et al (2019) Identifying EGFR mutations in lung adenocarcinoma by noninvasive imaging using radiomics features and random forest modeling. Eur Radiol 29:4742–4750. https://doi.org/10.1007/s00330-019-06024-y
11. Tu W, Sun G, Fan L et al (2019) Radiomics signature: a potential and incremental predictor for EGFR mutation status in NSCLC patients, comparison with CT morphology. Lung Cancer Amst Neth 132:28–35. https://doi.org/10.1016/j.lungcan.2019.03.025
12. Song J, Tian J, Zhang L et al (2019) Development and validation of a prognostic index for efficacy evaluation and prognosis of first-line chemotherapy in stage III–IV lung squamous cell carcinoma. Eur Radiol 29:2388–2398. https://doi.org/10.1007/s00330-018-5912-2
13. Park JE, Kim D, Kim HS et al (2019) Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol https://doi.org/10.1007/s00330-019-06360-z
14. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge
15. Wang X, Peng Y, Lu L, et al (2017) ChestX-Ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, pp. 3462–3471
16. Irvin J, Rajpurkar P, Ko M, et al (2019) CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. ArXiv190107031 Cs Eess
17. Depeursinge A, Vargas A, Platon A, Geissbuhler A, Poletti PA, Müller H (2012) Building a reference multimedia database for interstitial lung diseases. Comput Med Imaging Graph 36:227–238. https://doi.org/10.1016/j.compmedimag.2011.07.003
18. Setio AAA, Traverso A, de Bel T et al (2017) Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med Image Anal 42:1–13. https://doi.org/10.1016/j.media.2017.06.015
19. Kim HJ, Li G, Gjertson D et al (2008) Classification of parenchymal abnormality in scleroderma lung using a novel approach to denoise images collected via a multicenter study. Acad Radiol 15:1004–1016. https://doi.org/10.1016/j.acra.2008.03.011
20. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical image computing and computer-assisted intervention – MICCAI 2015. Springer International Publishing, Cham, pp 234–241
21. Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39:2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615
22. Vakalopoulou M, Chassagnon G, Bus N et al (2018) AtlasNet: Multi-deep non-linear elastic networks for multiorgan medical image segmentation. In: Medical image computing and computerassisted intervention − MICCAI 2018

23. Donahue J, Anne Hendricks L, Guadarrama S, et al (2015) Longterm recurrent convolutional networks for visual recognition and description. arXiv:1411.4389
24. Lee PQ, Guida A, Patterson S et al (2019) Model-free prostate cancer segmentation from dynamic contrast-enhanced MRI with recurrent convolutional networks: a feasibility study. Comput Med Imaging Graph 75:14–23. https://doi.org/10.1016/j.compmedimag.2019.04.006
25. Goodfellow I, Pouget-Abadie J, Mirza M et al (2014) Generative adversarial nets. In: Ghahramani Z,Welling M, Cortes C et al (eds) Advances in neural information processing systems 27. Curran Associates, Inc., pp 2672–2680
26. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302. https://doi.org/10.2307/1932409
27. Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: Lechevallier Y, Saporta G (eds) Proceedings of COMPSTAT'2010. Physica-Verlag HD, pp 177–186
28. Mollura DJ, Azene EM, Starikovsky A et al (2010) White paper report of the RAD-AID Conference on International Radiology for Developing Countries: identifying challenges, opportunities, and strategies for imaging services in the developing world. J Am Coll Radiol 7:495–500. https://doi.org/10.1016/j.jacr.2010.01.018
29. Kesselman A, Soroosh G, Mollura DJ, RAD-AID Conference Writing Group (2016) 2015 RAD-AID Conference on International Radiology for Developing Countries: the evolving global radiology landscape. J Am Coll Radiol 13:1139–1144. https://doi.org/10.1016/j.jacr.2016.03.028
30. Rajpurkar P, Irvin J, Lungren M, Langlotz C, Liang P (2019) Validating the CheXpert model on your own data in 30 minutes. In: github. https://rajpurkar.github.io/mlx/chexpert-validate/. Accessed 21 Oct 2019
31. Lakhani P, Sundaram B (2017) Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 284:574–582. https://doi.org/10.1148/radiol.2017162326
32. Jaeger S, Karargyris A, Candemir S et al (2014) Automatic tuberculosis screening using chest radiographs. IEEE Trans Med Imaging 33:233–245. https://doi.org/10.1109/TMI.2013.2284099
33. Melendez J, Hogeweg L, Sánchez CI et al (2018) Accuracy of an automated systemfor tuberculosis detection on chest radiographs in high-risk screening. Int J Tuberc Lung Dis 22:567–571. https://doi.org/10.5588/ijtld.17.0492
34. Bortsova G, Dubost F, Ørting S, et al (2018) Deep learning from label proportions for emphysema quantification. In: Frangi AF, Schnabel JA, Davatzikos C, et al (eds) Medical image computing and computer assisted intervention – MICCAI 2018. Springer International Publishing, pp 768–776
35. Walsh SLF, Calandriello L, Silva M, Sverzellati N (2018) Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. Lancet Respir Med 6:837–845. https://doi.org/10.1016/S2213-2600(18)30286-8
36. Wang S, Shi J, Ye Z, et al (2019) Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. Eur Respir J 53: https://doi.org/10.1183/13993003.00986-2018
37. Xu Y, Hosny A, Zeleznik R et al (2019) Deep learning predicts lung cancer treatment response fromserialmedical imaging. Clin Cancer Res 25:3266–3275. https://doi.org/10.1158/1078-0432.CCR-18-2495