Physica Medica 83 (2021) 9–24

Contents lists available at ScienceDirect

Physica Medica

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

Review paper

# AI applications to medical images: From machine learning to deep learning

Isabella Castiglioni [a,b,1], Leonardo Rundo [c,d,1], Marina Codari [e,1], Giovanni Di Leo [f], Christian Salvatore [g,h,*], Matteo Interlenghi [h], Francesca Gallivanone [b], Andrea Cozzi [i], Natascha Claudia D'Amico [j,k], Francesco Sardanelli [f,i]

[a] Department of Physics, Università degli Studi di Milano-Bicocca, Piazza della Scienza 3, 20126 Milano, Italy
[b] Institute of Biomedical Imaging and Physiology, National Research Council, Via Fratelli Cervi 93, 20090 Segrate, Italy
[c] Department of Radiology, Box 218, Cambridge Biomedical Campus, Cambridge CB2 0QQ, United Kingdom
[d] Cancer Research UK Cambridge Centre, University of Cambridge Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, United Kingdom
[e] Department of Radiology, Stanford University School of Medicine, Stanford University, 300 Pasteur Drive, Stanford, CA, USA
[f] Unit of Radiology, IRCCS Policlinico San Donato, Via Rodolfo Morandi 30, 20097 San Donato Milanese, Italy
[g] Scuola Universitaria Superiore IUSS Pavia, Piazza della Vittoria 15, 27100 Pavia, Italy
[h] DeepTrace Technologies S.r.l., Via Conservatorio 17, 20122 Milano, Italy
[i] Department of Biomedical Sciences for Health, Università degli Studi di Milano, Via Luigi Mangiagalli 31, 20133 Milano, Italy
[j] Department of Diagnostic Imaging and Stereotactic Radiosurgery, Centro Diagnostico Italiano S.p.A., Via Saint Bon 20, 20147 Milano, Italy
[k] Unit of Computer Systems and Bioinformatics, Department of Engineering, Università Campus Bio-Medico di Roma, Via Alvaro del Portillo 21, 00128 Roma, Italy

ARTICLE INFO

*Keywords:*
Artificial intelligence
Deep learning
Machine learning
Medical imaging
Radiomics

ABSTRACT

*Purpose:* Artificial intelligence (AI) models are playing an increasing role in biomedical research and healthcare services. This review focuses on challenges points to be clarified about how to develop AI applications as clinical decision support systems in the real-world context.
*Methods:* A narrative review has been performed including a critical assessment of articles published between 1989 and 2021 that guided challenging sections.
*Results:* We first illustrate the architectural characteristics of machine learning (ML)/radiomics and deep learning (DL) approaches. For ML/radiomics, the phases of feature selection and of training, validation, and testing are described. DL models are presented as multi-layered artificial/convolutional neural networks, allowing us to directly process images. The data curation section includes technical steps such as image labelling, image annotation (with segmentation as a crucial step in radiomics), data harmonization (enabling compensation for differences in imaging protocols that typically generate noise in non-AI imaging studies) and federated learning. Thereafter, we dedicate specific sections to: sample size calculation, considering multiple testing in AI approaches; procedures for data augmentation to work with limited and unbalanced datasets; and the interpretability of AI models (the so-called *black box* issue). Pros and cons for choosing ML *versus* DL to implement AI applications to medical imaging are finally presented in a synoptic way.
*Conclusions:* Biomedicine and healthcare systems are one of the most important fields for AI applications and medical imaging is probably the most suitable and promising domain. Clarification of specific challenging points facilitates the development of such systems and their translation to clinical practice.

## 1. Background

Artificial intelligence (AI) models are playing an increasing role in biomedical research and clinical practice, displaying their potential in several applications such as risk modelling and stratification, personalized screening, diagnosis (including classification of molecular disease subtypes), prediction of response to therapy, and prognosis [1]. These ground-breaking advances might yield a clinical impact by integrating

* Corresponding author at: DeepTrace Technologies S.r.l., Via Conservatorio 17, 20122 Milano, Italy.
*E-mail addresses:* isabella.castiglioni@unimib.it (I. Castiglioni), lr495@cam.ac.uk (L. Rundo), mcodari@stanford.edu (M. Codari), gianni.dileo77@gmail.com (G. Di Leo), salvatore@deeptracetech.com (C. Salvatore), interlenghi@deeptracetech.com (M. Interlenghi), francesca.gallivanone@ibfm.cnr.it (F. Gallivanone), andrea.cozzi1@unimi.it (A. Cozzi), nataschaclaudia.damico@cdi.it (N.C. D'Amico), francesco.sardanelli@unimi.it (F. Sardanelli).
[1] Isabella Castiglioni, Leonardo Rundo, and Marina Codari equally contributed to this paper.

https://doi.org/10.1016/j.ejmp.2021.02.006
Received 30 November 2020; Received in revised form 9 February 2021; Accepted 13 February 2021
Available online 1 March 2021

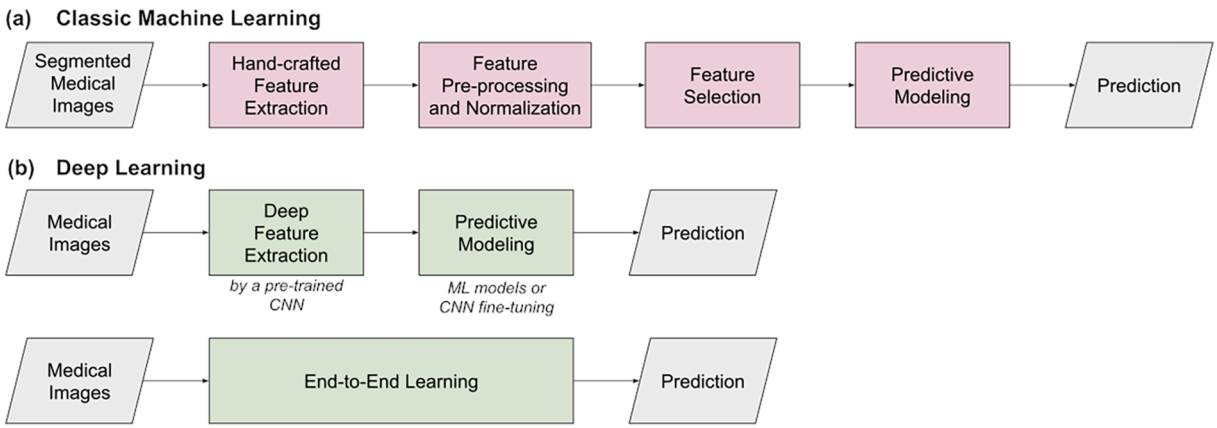


**Fig. 1.** Typical architecture and workflow of artificial intelligence systems for predictive modelling: a) classic machine learning, with the various processing steps involving hand-crafted features such as in radiomics; b) deep learning considering either deep medical image feature extraction or end-to-end learning.

multiple data flows from heterogeneous sources [2]. These sources include medical images, which constitute the largest part of patients' data (in particular for oncologic patients), but also disease risk factors, multiomics data, therapy procedures/regimens, and follow-up data. The effective integration of these sources in models leading to high-performance healthcare services will facilitate the convergence of human intelligence and AI sides [3,4]. All these research fields could greatly enhance the current trend toward precision medicine, resulting in more reliable and personalized approaches with high impact on diagnostic and therapeutic pathways [5]. This implies a paradigm shift from the definition of statistical and population-based outlooks to individual predictions [6], allowing for more effective preventive actions and therapy planning.

However, even though several guidelines have been already published on the development and usage of AI models [7–14], potential AI strategies are many and varied. Challenges and points to be better clarified do exist on "how to develop AI applications" as clinical decision support systems.

Consequently, we will focus herein on: the differences between the radiomic application domain, based on classic machine learning (ML) models and deep learning (DL) models, using multi-layered artificial neural networks, in particular convolutional neural networks (CNNs); specific AI issues for sample size calculation; procedures for *data augmentation* to work with limited and unbalanced datasets; data curation; interpretability of AI models (the "black box" issue). The section about data curation will include crucial technical steps such as *image labelling, image annotation* (with *segmentation* as a crucial step in radiomics), *data harmonization* (enabling compensation for differences in imaging protocols that typically generate noise in non-AI imaging studies) and *federated learning*.

Finally, together with some conclusive remarks, we provide pros and cons of choosing ML *versus* DL, along with some recommendations and references to existing software tools for AI developers and users, as well as essential take-home messages to the readers.

## 2. Methods and architectures for AI applications

Two different architectures and associated typical workflows can be implemented to develop AI applications in medical imaging (Fig. 1): (*i*) classic ML, exploiting hand-crafted features, namely radiomic features [15,16] extracted from segmented images; and (*ii*) DL, using deep feature extraction or end-to-end learning from images. However, ML and DL share general concepts, such as supervision and training, that must be clarified before considering specific aspects of the two approaches.

### 2.1. Supervised learning versus unsupervised learning

In AI-based classification systems, the most popular among *learning processes* is *supervised learning*, in which the training of the classification model is performed by presenting "labeled" *training data* (data samples coupled to their corresponding class or label of interest) to the learning system. The task of the learning system is then to find a relation that maps each input of the training set (the data) into an output (the label). In medicine, input data can include medical images or clinical data, while the output label can be, for example, the disease diagnosis, the patient condition (*e.g.*, the disease stage at a given follow-up time), the outcome after therapy (*e.g.*, recurrence, survival). Once such a relationship has been learned (*i.e.*, *training phase*), it can be used to classify new input data with unknown label into one of the classes of interest defined during the *training phase* [17].

In contrast to *supervised learning*, in *unsupervised learning* no *training data* is coupled to any pre-existing class or label of interest, possibly because of a lack of this information. The learning system is then fed with a set of *training data*, and its task is to search for undetected patterns that can separate these data into subsets of similar samples under given characteristics. Once these subsets and their characteristics have been detected and learned (*training phase*), new input data can be classified into one of the classes of interest that have been implicitly defined during the *learning process* itself (*i.e.*, the *testing phase*) [17].

Relevant examples of both supervised and unsupervised learning algorithms are given throughout this review. It must be noted that other approaches can also be used, such as *semi*-supervised learning, in which only part of the *training data* is labeled, making this approach a combination of supervised and unsupervised learning [17].

### 2.2. Training, validation, and testing

As introduced in the previous paragraph, the implementation of a classification model involves at least two phases, training and testing. The *training phase* is the one in which the learning of the classification model itself takes place. Data used in this phase are called *training data*, independently from the use of a *supervised* or *unsupervised* approach. To obtain a model with generalization abilities, *i.e.*, well-performing when applied to new data, the *training data* must be in a sufficiently large number and representative of the "general" population, *i.e.*, of the population on which the system will be tested and, finally, potentially applied in a clinical perspective [18].

The *testing phase* is the one in which the model learned during the *training phase* is used or tested on new samples. Data used in this phase are called *testing data*, and the performance of the model in correctly classifying these data is called *testing performance*. Of note, it is paramount that *none* of the samples included in the *training data* is used also

during the *testing phase*, as this would invalidate the *testing performance* [18].

To improve the learning performance, and when the available samples are enough in number, it is useful to introduce a third phase between the training and the *testing phases*, which is called *validation*. In this phase, the model parameters learned during the *training phase* are tuned and optimized to maximize a given metric (*e.g.*, its classification performance). Such parameters may include the number of variables used or their relative weight. Data used in this phase are called *validation data*, and the performance of the model in correctly classifying these data is called *validation performance*. It is important noting that the *testing performance* represents the final performance of the model, *i.e.*, the one that demonstrates the ability of the learned model to work on the general population [18].

### 2.3. Classic machine-learning models

According to Fig. 1a, the predictive modelling based on classic ML techniques starts from the extraction of large-scale hand-crafted features after regions of interest (ROIs) or volumes of interest (VOIs) have been either manually or (semi-)automatically delineated in the *image segmentation* process. This emerging research field, recently named "radiomics" [15], involves the extraction of mineable features from medical images to non-invasively characterize the *in vivo* phenotype of lesions or even simply of tissue portions (*e.g.*, the apparently normal tissue surrounding a tumor) [16], capturing the ROI/VOI characteristics by morphometric measurements (*i.e.*, size, shape, and diameter), as well as by measurements of tissue or function texture heterogeneity (including first-, second-, and higher-order statistical descriptors).

### 2.4. Radiomic application domain

Radiomic features are often not robust against medical-image acquisition parameters, such as spatial resolution (in-plane resolution and through-plane resolution, *i.e.*, slice thickness) [19,20] and image extraction settings (*e.g.*, quantization, resampling) [21,22]. Moreover, radiomic features can be dependent on the software package used to extract them [23].

These issues have been addressed by the *Image Biomarker Standardization Initiative* (IBSI) [24], which provided standardized definitions of radiomics features, computation, normalization, and nomenclature, also recommending how to implement the different steps of a radiomic workflow, including data conversion in standardized units, post-acquisition image processing, *image segmentation*, *data interpolation*, *resegmentation* (*i.e.*, procedure that involves only the pixels within a specified gray value range for radiomic feature calculation within the ROI/VOI), and intensity discretization. Their description is out of the purpose of the present review, being most of them well defined in the IBSI guideline [24]. Once features are computed and normalized, feature selection processes must be devised specifically for the radiomic domain, in order to define robust imaging biomarkers [25]. To this aim, the selection process should perform: (*i*) elimination of unreliable features (for instance *via* the intraclass correlation coefficient); (*ii*) elimination of not informative features based on zero and near-zero variance; and (*iii*) elimination of redundant features (*e.g.*, those which are highly correlated to each other). After these preprocessing steps, a further feature selection step aims at identifying the most relevant predictive features [26,27].

Importantly, all these techniques can deal with the "curse of dimensionality" and also reduce model *overfitting*, thus increasing the generalizability of the model. Feature selection methods can be subdivided into three classes: (*i*) *filter methods*, which leverage either statistical correlation or information theory-based metrics to assess the usefulness of a given feature subset; (*ii*) *wrapper methods*, which optimize the predictive model performance evaluating feature combinations using a search algorithm (*e.g.*, recursive feature elimination, sequential feature selection, metaheuristics); (*iii*) *embedded methods* allowing for feature selection as a part of the model, such as in the case of *least absolute shrinkage and selection operator* (LASSO) or *elastic net regularization methods* (*ElasticNet*). Among these methods, *wrapper methods* are powerful but computationally burdensome [28]. Indeed, they rely upon the evaluation of classification performance for obtaining the optimal feature subset: this search in the feature space is a non-deterministic polynomial-time hard (NP-hard) problem. Exhaustive search methods are computationally intensive and unfeasible for large-scale datasets, thus search methods and metaheuristics are typically used to find sub-optimal solutions in the search space [29]. Importantly, due to multiple statistical comparisons the repeated estimation of the accuracy employed in feature subset selection may cause *overfitting* in the feature subset space, thus hindering generalization abilities [30].

After obtaining a subset of reliable, nonredundant, and relevant features from these selection steps, the predictive model has still to be defined. This can be achieved by multivariable classification or regression methods according to the clinical question at hand [25,31], typically in supervised learning settings. The choice of either classification or regression approaches depends on whether the response (target) variable is categorical or continuous, respectively. It is worth noting that regression analysis can be used in classification tasks when binary or multinomial logistic models are employed. Alternatively, unsupervised clustering techniques can be used to identify intrinsic properties and patterns of input data (for instance, class grouping based on similarity metrics).

*Validation* of radiomic models represents another crucial phase. Although the performance of this step choice might depend on the available data quantity, *it is fundamental to avoid the use of the same data for both model training and testing*. Ideally, an independent dataset should be used as external test set. However, this is often not possible, and a single cohort must often be exploited for both model development and testing. Several strategies are available and can be used for this purpose.

One of the possible methods is the *hold-out* approach, which splits the whole dataset in one training set and one testing set (generally, 70% *versus* 30% or 80% *versus* 20%, respectively). This partitioning might be either random or based on a criterion (*e.g.*, temporal or center independence).

Other schemes, such as *cross-validation* (CV) strategies can be used. *Leave-one-out* CV (obtaining high variance and low bias) and *k-fold* CV are the most used schemes. *Leave-one-out* is often used when very few data are available to develop a ML model, but this method should be avoided because of its high variability, being based on a single observation. Of note, *k-fold* CV overcomes *leave-one-out* limits and improves the use of the available dataset compared to the *hold-out* method: the dataset is subdivided into *k* mutually exclusive folds of approximately equal size, allowing for a higher statistical validity [25]. The results for all the *k-fold* rounds are averaged, with a decreased dependency on the initial random split of the dataset, compared to the *hold-out* strategy.

The use of a *nested k-fold* scheme (with outer and an inner CV loops) is the most rigorous method allowing for model training independently from optimizing model hyperparameters [31]. Indeed, the hyperparameter selection by means of a *non-nested k-fold* scheme could yield a biased model, providing overoptimistic performance, since the selection of a model without *nested k-fold* CV implies using the same data to tune model hyperparameters and evaluate model performance, with potential *overfitting* on the *training data* and poor generalization ability.

Importantly, radiomic features can be integrated with additional information (*e.g.*, demographic data, risk factors, molecular data) to improve the predictive performance of the model. This integration is easiest with hand-crafted features, since supplemental data can be added in the ML model as additional features. In particular, multimodal imaging [32] and multiomics data [33] can be added to a model to better characterize the underlying pathophysiology of the analyzed image region. A Radiomics Quality Score [34] has been recently proposed to measure the quality of radiomic-based AI models, considering the

different steps occurring within the radiomic workflow. While there is still no consensus on its validity, it can usefully guide developers and users on verifying the completeness of the different features and tests to be implemented for providing an effective AI model.

### 2.5. Deep learning models

DL models (Fig. 1b) offer the opportunity to automatically extract imaging features to maximize model performance for the task at issue. DL is a specific subfield of ML that employs artificial neural networks, allowing to directly process raw data [35]. Indeed, deep neural networks enable the development of end-to-end predictive models by performing all the processing steps usually involved in the design of a classic ML model, including feature extraction and learning (see Fig. 1a).

Deep neural networks are representation-learning algorithms composed of a stack of processing layers with a finite number of nonlinear units (*i.e.*, artificial neurons). The first and the last layer of the network are defined as input and output layers, respectively, while all layers stacked between them are called *hidden layers*. The multi-layered structure of deep neural networks allows them to serve as nonlinear function approximators, able to learn different representations of the input data at multiple levels of abstraction [36]. Depending on the number of layers and units per layer, a DL model can easily reach millions of trainable parameters to be estimated during the training process. DL models are therefore prone to *overfitting*, especially when dealing with relatively small training sets, and are best applied to datasets of at least thousands of images [37].

Due to its ability to model very complex relationships within large datasets, DL has been largely applied in medical imaging and radiation oncology [38], with specific applications in the medical imaging domain including both large and small image datasets, although with different implications.

Among different neural network architectures, CNNs are the most used for medical-image processing tasks. These networks are characterized by the presence of convolutional layers between the layers of neurons, convolving an input image with a given kernel function. In CNNs, different convolution layers can be implemented according to the application purpose, since the weights of convolutional layers being learned during training can extract imaging features tailored to the investigated task. Compared to fully connected neural networks, in CNNs the same kernel parameters are applied to the entire image, thus reducing the overall number of trainable parameters and making the training process more efficient. Depending on input and output data dimensionality, one-, two, or three-dimensional convolutional kernels can be employed.

Pooling layers are another key component of CNNs architecture: they reduce feature map resolution to introduce translational invariance to minor image distortions. Moreover, the combination of convolutional and pooling layers allows for learning spatial hierarchies among feature patterns [39].

The stack of linear (convolution) and nonlinear (activation) processing layers operates as a feature extractor, progressively increasing the level of abstraction, invariance, and discriminative power across layers [40]. After this processing, these features are then combined by either a series of fully connected layers or by other classic ML algorithms that perform the learning task (Fig. 1b).

Convolutional, pooling, and activation layers are not the only possible components of CNN architectures. Due to the modular structure of CNNs, several architectures have been proposed combining CNN with other types of neural networks. *End-to-end CNN architectures* that directly map images to a target class have been used to perform image classification tasks for both screening and diagnosis purposes. In particular, several CNN architectures originally trained on large natural image datasets, such as *ImageNet*, have been employed for medical image classification by fine-tuning pretrained layers to address data sparsity issues [41]. Introduced in 2015, the *U-Net architecture* is still one of the most used CNN architectures for medical *image segmentation*. The base *U-Net architecture* is composed of symmetrical encoder and decoder paths connected using skip connections. Originally proposed to process two-dimensional images, it has been modified to obtain voxel-wise *segmentation* from three-dimensional images [42,43]. Then, to further improve network performances, several variants of this network have been developed by adding residual, attention, or *DenseNet* blocks to train deeper networks, select salient features, and solve gradient vanishing issues, respectively [44]. The above-mentioned architectures represent a brief introduction to the broad spectrum of available architectures: a detailed taxonomy of CNN architectures is out of the purpose of the present article but can be found in the recent review by Khan et al. [45].

*Recurrent neural networks* (RNNs) have also been combined with CNNs to extract spatial–temporal features from imaging data series. These networks allow for processing new data (*e.g.*, image series of any size) while being aware of previous inputs and outputs by sharing node weights across time. However, model complexity is directly proportional to the size of input data, making RNNs difficult to train and prone to *overfitting*. To address vanishing/exploding gradient issues and allow to memorize long term information, gated recurrent units and Long Short-Term Memory (LSTM) units have been introduced [46].

*Autoencoders* also play a pivotal role among *unsupervised* DL architectures, learning in an unsupervised way how to reproduce the input data. In these networks, the use of progressively smaller *hidden layers* in the encoder path, regularization, and sparsity constraints, allow to learn a lower-dimensional representation of the data, thus preventing the network from learning the identity transformation (*i.e.*, the trivial solution) [38]. More recently, *generative adversarial networks* (GANs) [47] have been widely used for medical image processing due to their ability to model data distribution and generate realistic datasets. GANs involve the interaction of two *adversarial networks*, where a network generates new realistic data by learning data distribution from training samples, and the other network discriminates between fake and real data. The interaction of these *adversarial networks* improves overall GAN performance and generates realistic image data (*i.e.*, *adversarial training* framework). Despite their innovative design, these networks are usually challenging to train due to vanishing/exploding gradient issues and are prone to generating new data with similar appearance (*i.e.*, model collapse) [48].

After selecting the proper network architecture, the hyperparameter tuning represents a non-trivial step. Designing the correct architecture is challenging, since several structural hyperparameters, such as the number of layers/neuronal units, the receptive field size (the region in the input space that a particular CNN's feature is looking at), and the activation functions can strongly affect model performance [49].

During learning, the network parameters are optimized to solve a specific task. To this aim, a backpropagation algorithm of the error adjusts the parameters of the network to minimize a loss function that represent the cost function of the network. The adjustment is based on the change of gradient of the loss function with respect to network parameters. To improve this process, several optimizers have been proposed. Along with the stochastic gradient descent, most of them employ adaptive learning rates to improve global minimum detection in complex optimization problems [50]. Moreover, input image normalization, as well as the use of batch normalization layers standardizing the automatically extracted deep features, have shown to help training convergence and prevent covariate shift [51].

The depth of the network should increase with the complexity of the investigated task. However, very deep neural networks are prone to the problem of vanishing/exploding the gradients, a problem that effectively prevents the weights from changing values during training, which may cause very long training time or failure to converge, respectively. The use of *Rectified Linear Unit* (ReLU) activation function, proper initialization techniques and skip connections may partially mitigate this issue [52]. Since excessive increase in model complexity may also result in *overfitting*, several regularization techniques can be used to

improve model generalizability, such as $L_1$ and $L_2$ regularization, batch-normalization, dropout, early stopping, and *data augmentation* techniques. These techniques can be combined to take advantage of the complementary effects of different approaches, as detailed in a comprehensive overview [53] of the most frequently adopted regularization techniques and of their effects on DL model performance.

Regarding the design choices, the "no free lunch" theory demonstrates that each model requires a specific hyperparameter setting to maximize its performance on a specific task [54]. Therefore, hyperparameter tuning represents a utterly needed albeit challenging and time-consuming step, which requires the continuous evaluation of model prediction error on *training* and *validation datasets* to find out the acceptable tradeoff between *overfitting* and *underfitting*. To find the best hyperparameter set, several approaches can be used. Traditional approaches range from exhaustive to random and multistep hyperparameter search, while more recently proposed approaches include automatic hyperparameter optimization algorithms, that reduce the burden of hyperparameter tuning on the model design process. In this scenario, reinforcement learning [55] and metaheuristic algorithms [56] represent promising alternatives to trial-and-error approaches. Still, the evaluation of DL model performance must mandatorily be done on the test set, which represents the only independent and external data set that can ensure model generalizability.

### 2.6. Deep learning in the medical image application domain

Training and evaluating deep neural networks in medical images can be more challenging than radiomic analysis with ML, mainly for the frequent lack of availability of a sufficient number of well-labelled medical image data. To solve this issue, *image augmentation* and *transfer learning* techniques can be used [57]. In this light, GANs can be used to generate synthetic additional training instances [58].

Alternatively, *deep transfer learning* techniques, which relax the hypothesis that training and *testing data* comes from the same probability distribution, allow avoiding training DL models from scratch. *Deep transfer learning* techniques have been classified into four categories: instances-, mapping-, network-, and *adversarial*-based, as detailed by Tan et al. [59].

Other ways to address the lack of properly annotated data is to use a *semi-supervised* or a *weakly-supervised* approach. In *fully-supervised* learning, labeled instances are used to train, validate, and test a DL model, while *weakly-supervised* approaches allow the exploitation of partially- or weakly-labelled data. Such strategies involve the use of partially labelled datasets (*incomplete supervision*), coarse-grained labelled datasets (*inexact supervision*), and datasets with not-only ground-truth labels (*inaccurate supervision*) [60]. Finally, recent advancements in DL research highlight the potential of *self-supervised* or *unsupervised* pre-training strategies: in *self-supervised* approaches, labels are automatically retrieved from data [61], while in *unsupervised* approaches imaging features are extracted without labels [62–64].

Both for DL and for ML, in the growing framework of personalized and precision medicine, another important challenge is the integration of different data modality features into a single model. This issue is particularly relevant when imaging and clinical data must be integrated with other omics data in a single DL model. In this light, the review article published by Li et al. [65] offers a comprehensive survey of available integration strategies, starting from ML but also covering multimodal DL integration strategies.

Along with *adversarial learning* applications to *data augmentation* and *transfer learning*, *adversarial attacks* are worth to be mentioned. The preparation of adversarial samples – by applying small modifications to the medical imaging samples that are close to the decision boundaries learned by a classifier [66] – might affect DL-based computer-assisted diagnosis systems [67], but also radiomics-based models [68]. Indeed, small changes to the pixel data might suitably change the values of some radiomic features that influence the downstream analyses. This problem cannot be ignored in reliable computer-assisted diagnosis systems that have to be employed in the clinical practice.

Considering the ever-increasing expansion of AI-focused literature in medical imaging, a guide for the development of reliable DL models for medical image analysis (Checklist for Artificial Intelligence in Medical Imaging, CLAIM), including recommendations on AI models generalizability and reproducibility, has been recently proposed [69].

## 3. Data quantity in AI applications

### 3.1. Sample size

In a typical AI classification task in oncologic imaging, an AI model aims to distinguish benign from malignant lesions using an imaging biomarker or radiomic features potentially associated with lesion characterization. In this case, distributions of malignant lesions are expected to be different from those of benign lesions, classically substantiated by a *p*-value [70].

AI applications usually involve hundreds or even thousands of statistical hypothesis tests. This largely increases the probability of false discoveries, *i.e.*, associations/correlations that lead to a statistically significant *p*-value, historically set at $< 0.05$, but are not actually true. For example, if one thousand statistical tests are performed at an alpha (type I) error of 0.05, 50 false discoveries would appear, on average. To mitigate this phenomenon, much lower significance thresholds could be adopted in these peculiar contexts [70]. The false discovery rate is intimately connected to the sample size: the larger the latter, the lower the former, and *vice versa*. Thus, the sample size is the major determinant of an AI model performance: small sizes of the training and test sets are sources of bias and contribute to the variance of a model performance [71,72].

In classical statistics, methods for sample size determination are well established for the several possible contexts (study design, outcomes, null hypothesis, etc.) that substantially build around the formula:

$$n = \left(\frac{Z\sigma}{E}\right)^2$$

This equation provides the size $n$ for a desired error rate ($E$) and variance ($\sigma$); $Z$ is the $Z$-distribution value for a given level of confidence. However, the above formula does not take into consideration any of the peculiar characteristics of AI modelling. Indeed, methods for calculating the required sample size in AI applications remain unclear and many researchers simply follow the Widrow-Hoff learning rule [73], *the empiric rule for multivariate analysis that suggests ten data (patients) for every imaging feature that will be used in the model.* This rule, however, may come up with a too small or too large sample size, depending on the context.

More analytic approaches for sample size calculation in the medical imaging field have been recently assessed in a systematic review by Balki et al. [74], where different methods were categorized into *model-based* (*i.e.*, based on algorithm characteristics) and *curve-fitting* methods (*i.e.*, empirically evaluating model performance at selected sample sizes). *Model-based* methods are built on the assumption that training and test samples are chosen from the same distribution. One was postulated by Baum and Haussler [60] for single hidden-layer feedforward neural networks with $k$ units and $d$ weights. This method predicts that for a classification error $\varepsilon$ ($0 < \varepsilon < 1/8$), a network trained on $m$ samples with the fraction $1\text{-}\varepsilon/2$ of the samples correctly classified, would approach a classification accuracy of $1\text{-}\varepsilon$ on an unseen test set, with the condition that $m \geq O(d/\varepsilon \cdot \log 2(k/\varepsilon))$. Another *model-based* method was proposed by Haykin [61] for whom generalization is valid if the condition $m = O((d + k)/\varepsilon)$ is satisfied. This method is similar to the Widrow-Hoff rule and, in practice, $m \approx (d/\varepsilon)$ [75].

Learning *curve-fitting* methods aim at modeling the relationship between training set size and classification accuracy using an inverse

power law function. Fukunaga and Hayes [76] proposed to empirically obtain the area under the curve at receiving operator characteristic through performance-testing procedures and to plot it against their respective $1/N_{train}$ ($N_{train}$ = number of training images): the performance at higher sample sizes is extrapolated by linear regression as $N_{train}$ tends to infinity. Although these pseudo methods provide *post-hoc* sample size estimates, an empirical approach has the advantage of accurately modelling performance for a specific task, avoiding assumptions on distributions.

Another promising method is based on the Vapnik–Chervonenkis (VC) dimension that simply estimates the power of a classification AI algorithm [77]. The sample size estimated through this method is based on the following equation:

$$Pr\left(E_{Test} \le E_{Training} + \sqrt{\frac{1}{N}\left[D\left(log\left(\frac{2N}{D}\right)+1\right) - log\left(\frac{\eta}{4}\right)\right]}\right)$$

which gives a probabilistic upper bound of test error ($E_{Test}$) generated by an algorithm upon the training error ($E_{Training}$); $D$ is the algorithm's VC dimension, $N$ is the sample size, and $0 \le \eta \le 1$. The sample size for a prespecified test error and a known training error may be calculated solving the above formula for $N$. Of course, the lower the desired test error, the larger the $N$.

In order to explore the performance of the AI system, to assess how much it is statistically different from chance, and to exclude the presence of false discoveries, it is desirable to apply a permutation test (which consists in training, validating, and testing an AI system using randomly-permuted gold-standard labels instead of original true labels) at the end of the learning-and-classification process [78]. This is particularly useful when 1) the size of the *training/testing dataset* is not high, 2) the training and/or testing subsets are not representative of the general population, or 3) the training of the AI model is heavily affected by confounding/noisy variables in the *training/testing dataset*. In all these cases, the resulting AI system may be more performant than expected.

### 3.2. Data augmentation

*Data augmentation* is a data-space solution to the problem of small data sets. Several techniques that enhance the size and variety of *training datasets* can be implemented, falling into two general categories: *data warping* and *oversampling* [79].

*Data warping* transform original images preserving their labels. Typical transformations include geometric and color transformations, cropping, noise injection [80], filtering [81], as well as mixing images together by averaging their pixel values [82] or generating images based on Monte Carlo simulated projections [83,84]. *Data oversampling* create synthetic instances in the space of the features (see Section 3.3.).

A completely different solution for *data augmentation* is *adversarial training*, *i.e.,* using two or more networks with contrasting objectives encoded in their loss functions. Li et al. [85] experimented with *adversarial training* and found improved model performance on original *testing data* enriched with adversarial instances. Following similar principles, the aforementioned GANs [47] create artificial instances from a dataset such that they retain similar characteristics to the original set. The use of GANs in medical imaging has been well documented in a survey by Yi et al. [86] and in further published studies applied to computed tomography (CT) [87], magnetic resonance imaging (MRI) [88], and X-ray [89] images. Using GAN-based *data augmentation*, an improvement in classification performance by 4–8% has been reported [90]. However, *data warping*, *oversampling* and *adversarial training* can be also used in combination, since they are not mutually exclusive: traditional hand-crafted *data warping* techniques can be used in combination with GANs.

There is still no consensus on the final augmented dataset size to achieve for improving an AI model. Over-extensive augmented data can cause *overfitting* of the AI model even worse than before *augmentation*. Thus, a good method is to monitor *overfitting* during *incremental augmentation* and define the maximum level of *data augmentation* on the maximum training accuracy and minimum loss.

### 3.3. Imbalance learning in AI applications

Another very common issue in biomedical AI applications linked to data quantity occurs when data is distributed over different classes with a large degree of sample size differences among them. This issue is typically due to a lower prevalence of some classes. In developing AI application, this issue is known as *imbalance learning*.

#### 3.3.1. Data resampling

Different *data resampling* approaches can be used to mitigate this problem, namely *undersampling* and *oversampling* methods. Both types of approaches resize the *training dataset* to achieve a more balanced class distribution, matching the size of other class(es): in *undersampling*, a subset of instances is samples from the majority class, while *oversampling* generates artificial samples to supplement the minority class. In case of *imbalance learning* in a multiclass framework, both *undersampling* and *oversampling* are usually applied in a pairwise scheme among the classes.

When the number of samples per class leads to discard *undersampling* methods, the following popular *oversampling* approaches can be used [91].

*Synthetic minority over-sampling technique* (SMOTE) is a standard benchmark for learning from imbalanced data: synthetic samples created in the feature space along segments joining any or all of the $k$ minority class nearest neighbors, randomly chosen (for example $k = 2$) [91]. Synthetic samples are generated by: (*i*) computing the difference between the feature vector under consideration and its nearest neighbor; (*ii*) multiplying the difference by a random number in [0, 1]; and (*iii*) adding this quantity to the feature vector under consideration. This corresponds to the selection of a random point along the segment between two specific features [92]. Of note, this approach proved successful in several domains, also inspiring other approaches to counteract class imbalance and significantly fostering new *semi-supervised* learning paradigms, such as multilabel classification and incremental learning [93].

*Borderline SMOTE* is based on the original SMOTE implementation but, rather than generating new samples from all minority class samples, first selects all borderline minority samples and, considering this selection, subsequently generates synthetic samples [94]. For every sample in the minority class, *borderline SMOTE* calculates the $m$-nearest neighbors from the whole training set, also determining the number of majority samples among these nearest neighbors. If the number of its majority nearest neighbors is larger than the number of its minority ones, the sample is considered to be easily misclassified and put into a set referred to as "danger" [91]. Otherwise, it is considered to be safe or to be noise, therefore exiting the *oversampling* procedure. Minority samples in the danger set represent borderline data of the minority class, and synthetic samples are finally generated applying the SMOTE algorithm [91]. This is the first implementation, referred to as *borderline SMOTE1*, while a second implementation (*borderline SMOTE2*) generates synthetic samples from each sample in the danger set, considering not only its nearest neighbors in the minority class, as SMOTE does, but also from its nearest majority neighbor [91].

The *adaptive synthetic sampling approach* (ADASYN) represents another improvement of SMOTE, essentially using a weighted distribution for different minority class samples, according to their level of difficulty in learning, as described by Haibo et al. [95]. In ADASYN, more synthetic data are generated for minority class samples that are harder to learn compared to easier-to-learn minority samples [91]. As a result, ADASYN reduces the bias introduced by the class imbalance, and it shifts the classification decision boundary toward the difficult samples [91,95].

#### 3.3.2. Ensemble learning

This approach employs an ensemble of learners, with each composing classifier ($C_i$) being trained both on a subset of the majority class and on a subset of the minority class, still accounting, however, for a large portion of the minority class samples [96–98]. Then, decisions taken by all $C_i$ on the test sample are combined to obtain a final output according to a given rule, such as majority voting [91]. The rationale of *ensemble learning* lies in the observation that an ensemble of classifiers generally yields better performance than those obtained by individual models [99,100], especially for generalization purposes. Furthermore, base classifiers $C_i$ are now trained on more balanced subproblems than the original one, also having the desired property of containing samples representing different aspects of the original set $N$ [91]. Three popular approaches can be described.

The *balanced bagging classifier* builds several learners on different randomly selected subset of data, balancing each subset of data by *undersampling* the majority class so that the number of selected samples matches the number of samples extracted from the minority class [91].

*Forest of randomized trees* is a variation of the original *random forest* method that builds an ensemble of trees induced from balanced and down-sampled data. First, for each iteration in *random forest*, a bootstrap sample is drawn from the minority class and, randomly, the same number of cases is drawn with replacement from the majority class. Second, a *classification and regression trees* (CARTs) classification process is started from the data to maximum size, without pruning. At each node, instead of searching through all variables for the optimal split, a subset of randomly selected variables is considered. Third, the two previous steps are repeated and, after training, the final decision is obtained by majority voting on each tree decision [101].

*XGBoost* is an optimized, scalable, portable and distributed implementation of *gradient boosting* [102], where the ensemble of trees is a CART. Deriving from a regularized objectivation of *gradient boosting*, this approach has recently gained much popularity as the algorithm used by several teams to win ML competitions [91]. Compared to *decision trees*, the leaf nodes in CARTs store a real-valued score rather than binary decision values. In this way, richer interpretations can be attained.

## 4. Data curation

Despite their differences, ML and DL share several challenges. As mentioned before, data collection and curation represent fundamental steps of data-driven model development [57].Especially in the case of medical images, the "garbage-in, garbage-out" principle remains valid [103]: the quality of the pool of images provided as input to any processing algorithm determines the reliability of the results, even for AI applications. The quality check of the images used to infer new knowledge is a particularly critical point, considering also that AI applications need to work on large sample sizes (high data quantity) with medical images often acquired in multicenter studies (high data heterogeneity due to different equipment, imaging and clinical protocols, etc.).

The assumption that AI only needs to be fed with random data collected and combined on a huge scale can gravely backfire. Incorrect datasets can come in many forms, ranging from factually incorrect information to knowledge gaps, incorrect conclusions and, finally, wrong clinical indications: an uncurated dataset can be biased, inaccurate, unreliable, partially represented, error-ridden, or ambiguous. Using uncurated raw datasets was "*found to decrease the feature quality when evaluated on a transfer task*" [64].

### 4.1. Data labelling and annotations

*Data labelling* aims at ensuring that the data set works for the model target. For example, an AI model based on medical images developed to predict different prognostic outcomes will need data labelled as images of good or poor prognosis. This step links the images to ground-truth information and implies to collect knowledge from histopathology on needle biopsy or surgical specimens, from laboratory results, from patients' clinical records, or even from patients' follow-up. Such knowledge can also represent ground-truth for other tasks such as AI applications for automatic first-level screening reading (as in screening mammography [104]), when the AI tool provides an immediate dichotomic classification into negative cases or recall cases, the former to be sent to the next screening round, the latter to be recalled for a suspicious lesion assessment..

In general, images can be labelled in different ways including structured labels, *image annotations*, and *image segmentations* [105,106]. While structured reporting of diagnostic imaging, as suggested by various guidelines, would strongly reduce the effort needed to extract labels, most clinical reports still unfortunately remain composed of free text [107]. As a result, most centers looking at using retrospective data have to manage large volumes of medical images associated to narrative reports, whose analysis requires huge efforts. Even though DL itself has been proposed for translating free text into structured reports, for example in CT pulmonary angiography [108], retrospective report-based *image labelling* is often done manually.

For instance, *image annotations* of radiological diagnoses can be done by using radiological reporting categories attributed to the lesion(s), such as the categories defined by the Breast Imaging Reporting and Data System, BI-RADS [109] or by the Prostate Imaging Reporting and Data System, PI-RADS [110]. *Image annotations* are mandatory also when informing the algorithm on the location of the lesion(s) or other specific tissue regions.

Groups of scientists have been employed in the past to perform *data labelling* and *annotation*, including *image segmentation* [111,112]. During competitions, data labelled by consensus are provided by the organizers to participants [113], such as in the The Crowds Cure Cancer project [114], where hundreds of participants who attended the 2017 and 2018 meetings of the Radiological Society of North America were involved in *image labelling* tasks for the Cancer Imaging Archive (https://www.cancerimagingarchive.net/).

Another fundamental aspect whose impact is most often underestimated in AI applications is *image segmentation*. While DL approaches do not always require the preliminary identification of ROIs or VOIs to extract imaging features for model training, *this step is mandatory for radiomics*: the more accurate the definition of the area/volume to be characterized, the more the extracted quantitative features entering the ML model will reflect the biological characteristics of the lesion or tissue.

Years of research on *image segmentation* algorithms have highlighted the aspects to be considered especially when using handcrafted imaging features coming from different imaging modalities and techniques, with or without administration of contrast agents or radiopharmaceuticals. *Image segmentation* methods are influenced by the characteristics of the lesion and of the images under consideration, especially in the case of hybrid and multimodal imaging [115].

Since the definition of ROIs or VOIs quantitatively impacts the radiomic characteristics [115,116] the results of radiomic analyses obtained using different *segmentation* methods can widely differ. To date, there is no consensus on the approaches to be used for *image segmentation* in radiomics studies. The IBSI standardization initiative [24] proposes, as a good compromise, the use of semiautomatic algorithms, including the use of fully automatic methods followed by manual adjustments by the operator, speeding up the process but still allowing for human correction. Notably, this has consequences on the stability of radiomic features: different *segmentation*/adjusting methods as well as different operators can cause variations in the computed radiomic features.

*A strategy against feature instability is to select the radiomics features that are statistically stable*, by applying different *segmentation* methods or asking different operators to segment images, in repeated (*test–retest*) studies, either on patients [16] or on anthropomorphic phantoms [115,117]. *Another strategy is to apply moderate random variations on segmented ROIs/VOIs provided by a single operator*. This process generates

different *segmentation* results, as if they were obtained by different *segmentation* methods or operators, without the need for other annotators and without the need to develop and implement alternative *segmentation* methods.

*Image segmentation* is required in DL models for image classification and object detection. Various *image annotation* techniques can be used with the help of ML algorithms providing bounding boxes, polygon *annotations*, cuboid *annotations*, and contours circumscribing a target on the image. This process, known as *semantic segmentation*, can enable in-depth detection of targeted objects associated with a disease, segmented in a single class and in a single process.

### 4.2. Data harmonization

Even when considering a single imaging modality, medical images can be acquired using different scanners or with the same scanner but with different clinical protocols and/or acquisition/reconstruction technical parameters. This leads to variable spatial resolution, contrast-to-noise ratio, and temporal resolution in dynamic contrast-enhanced studies. The impact of these variations on the robustness of radiomic analysis has already been reported. Meyer et al. [118] showed that more than 80% of the radiomic features extracted from CT images were found not reproducible considering different settings of image reconstruction and radiation dose. Similarly, both phantom [115] and clinical studies [119] demonstrated that radiomic features from positron emission tomography (PET) are strongly influenced by reconstruction settings, while magnet field strength, type of scanner, and acquisition parameters have a similar impact on MRI studies [120–122]. Studies employing DL seem to be less limited by this problem, but we are still far from clear understanding of whether this is an effect of feature extraction techniques [123] or, as already mentioned, of the intrinsically higher sample size that reinforces biomedical data robustness and reproducibility [103]. Recently, *data harmonization* techniques have been developed to compensate for the aforementioned variations [124]: such methods normalize the statistical distributions of the same features when obtained from different systems, preserving the information content of images [125].

### 4.3. Image intensity normalization, denoising and artifacts corrections

Another problem that can have a non-negligible impact on image-analysis algorithms in both radiomics and DL approaches concerns the use of arbitrary units to measure the signal, typically in MRI. In fact, PET and CT images have units of measure based on well-defined physical processes so that the signal has a quantifiable physiological meaning: being calibrated according to agreed standards, a statistically significant variation in the signal can be interpreted as a real one. Conversely, MRI provides images whose signal is expressed in arbitrary units, hindering the comparison of images captured not only in a population study but even in longitudinal studies in the same subject. Interesting exceptions to this general paradigm are represented by apparent diffusion coefficient (ADC) maps derived from diffusion-weighted sequences [126], T1 and T2 mapping [127], as well as by MRI fingerprinting [128]. Hence, in MRI, *denoising* and *intensity normalization* procedures are required before extracting quantitative biomarkers from the images for use in AI applications. Different normalization methods have been described: scaling and shifting of whole imaging values to fixed intensity range [129]; normalizing to whole image mean and standard deviation [130]; normalizing to a biologically comparable reference tissue region [131]; and adjusting imaging histogram to a reference one [132]. Even if no definitive conclusions were obtained, several studies [22,133] have shown how these multiparametric MRI image corrections can impact the value of radiomic features by improving the performance of AI applications.

Besides *image intensity normalization*, MRI images can be corrected for noise and artifacts. A wide range of *denoising* methods have been proposed: bilateral filtering methods [134]; nonlocal means filtering methods [135]; block matching; and three-dimensional filtering methods [136] or global filters [137]. Bias field correction (BFC) refers to corrections allowing to compensate for magnetic field inhomogeneities, such as the N4ITK algorithm currently employed in most radiomics studies to perform BFC [138]. However, no large studies specifically addressed the impact of such correction on the AI performance.

Even images that provide quantitative parameters, such as PET

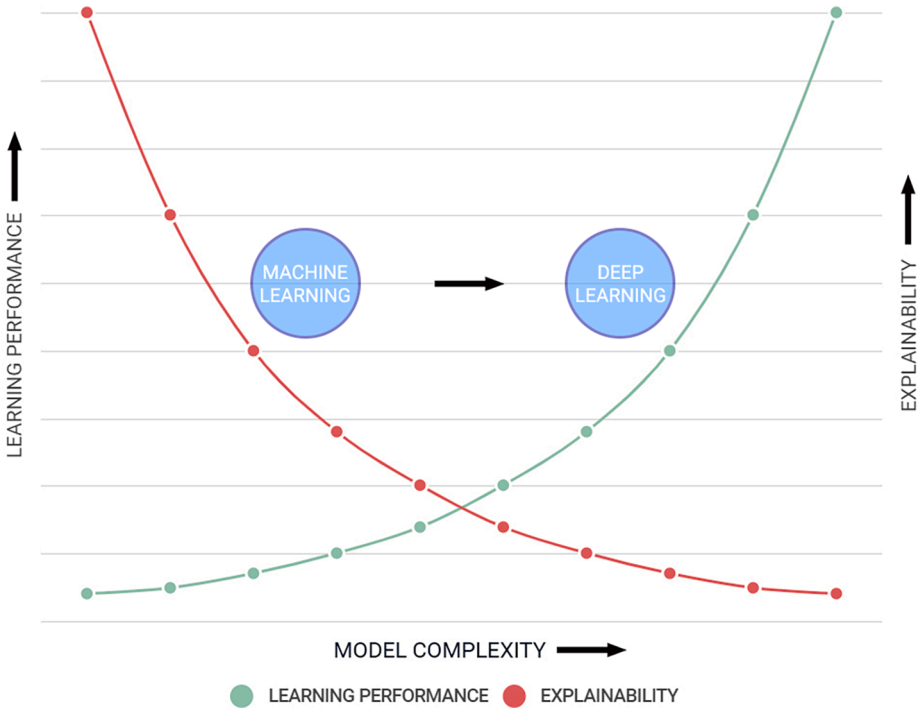


**Fig. 2.** Learning performance and explainability of an artificial intelligence system as a function of model complexity.

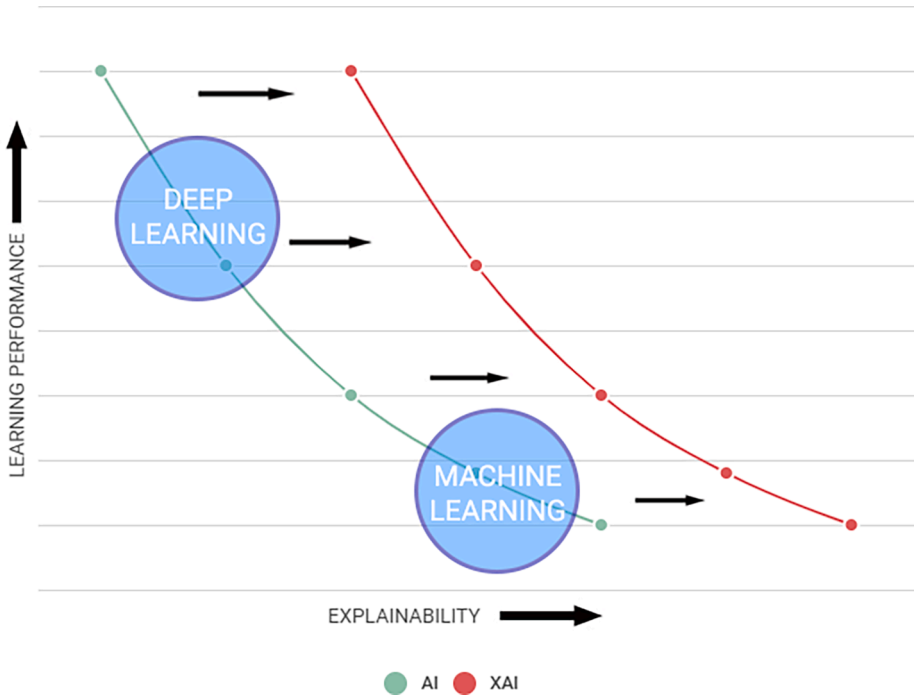


**Fig. 3.** Learning performance as a function of explainability for artificial intelligence (AI) *versus* explainable artificial intelligence (XAI).

standardized uptake value (SUV) units or MRI ADC maps, are often subject to a wide spectrum of physical effects that generate possible artifacts. This is a relevant issue for ADC maps [139,140], making this "exception" only a relative one. In these cases, radiomics or DL studies will benefit from correction methods impacting on the entire image, such as those methods compensating physical effects during image reconstruction [141]. As pointed out by Litjens et al. [38], image processing algorithms such as *intensity normalization* and *denoising* have not yet been widely used in the context of DL algorithms, probably due to the large number of images used in DL studies that acts as a compensating factor. Still, some studies suggest that these corrections may help improving the performance even of DL models [38,103], and we expect that their use will increase in the future.

### 4.4. Applicability of federated learning

Along with careful data *annotation* and *harmonization*, the potential applicability of *federated learning* deserves to be discussed. Indeed, large-scale data collections do not introduce only logistic problems due to the exchange of massive datasets across different institutions, but multicentric and international AI-powered studies have also to deal with strict and rigorous regulations regarding ethical and legal aspects of patient data exchange [142,143]. As a matter of fact, in medical imaging [144], the storage and transfer of the scans is facilitated by the Digital Imaging and COmmunication in Medicine (DICOM) standard [145,146]. The traditional method of training AI models involves setting up servers where models are trained on data, often using a cloud-based computing platform. However, an alternative way of model creation has arisen, called *federated learning*, which brings ML to the data source, rather than bringing the data to the model.

In *federated learning*, trained consensus models are developed exploiting data collected by different institutions without the need for sharing them and maintaining patient privacy. By implementing a decentralized data model and performing computations either by aggregation servers or *via* peer-to-peer systems, this approach offers controlled and secure access to large, heterogeneous, and curated multicentric datasets for both development and evaluation purposes [144,147]. However, the potential of *federated learning* requires a huge effort from involved participants to ensure high standardization and reliability at each step of the model development process, from patient enrollment to model evaluation, especially in terms of model generalizability. Regarding the realization of *federated learning* infrastructures, each partner has to assure valuable high-performance computing (HPC) resources in terms of hardware, software and network bandwidth [147]. As a virtuous side-effect, this need could lead to the substantial strengthening of the HPC resources in healthcare environments.

## 5. Interpretability of AI applications

As presented so far, AI applications to medical images have shown continuous improvements, both in the implementation of new technologies for learning, automatic classification, and prediction, as well as considering the intrinsic performance obtained in various fields. However, the increase in complexity of techniques and developed models corresponds to an increased difficulty in understanding the underlying learning and classification processes [148]. A typical example of this behavior (Fig. 2) can be seen in the translation from ML techniques to DL architectures.

Recently, the need to make AI reasoning transparent and intelligible to human readers has strongly emerged, with the aim to see, study, and understand how inputs are mathematically mapped into outputs [149,150] and to clarify the patterns within the inner mechanisms of AI systems. An AI system able to describe its behavior – or the behavior of AI-controlled entities – is called *eXplainable AI* (XAI), a term first introduced by Van Lent et al. in 2004 for simulation-game applications [151]. The term "explainability" can also be expressed as "understandability" [152], "comprehensibility" [153], "intelligibility" [154], or "interpretability" [149]: it is however clear from these definitions that the development of XAI systems should not in any way affect the classification/prediction performance of the models, but only their explainability, as shown in Fig. 3.

The need for XAI is particularly pronounced in those fields that require high transparency, as is for the biomedical field, where the reliability of AI systems in decision making should be strongly documented [155] if their use is proposed to support clinicians and patients in their decisions. Other important issues regard the clinical

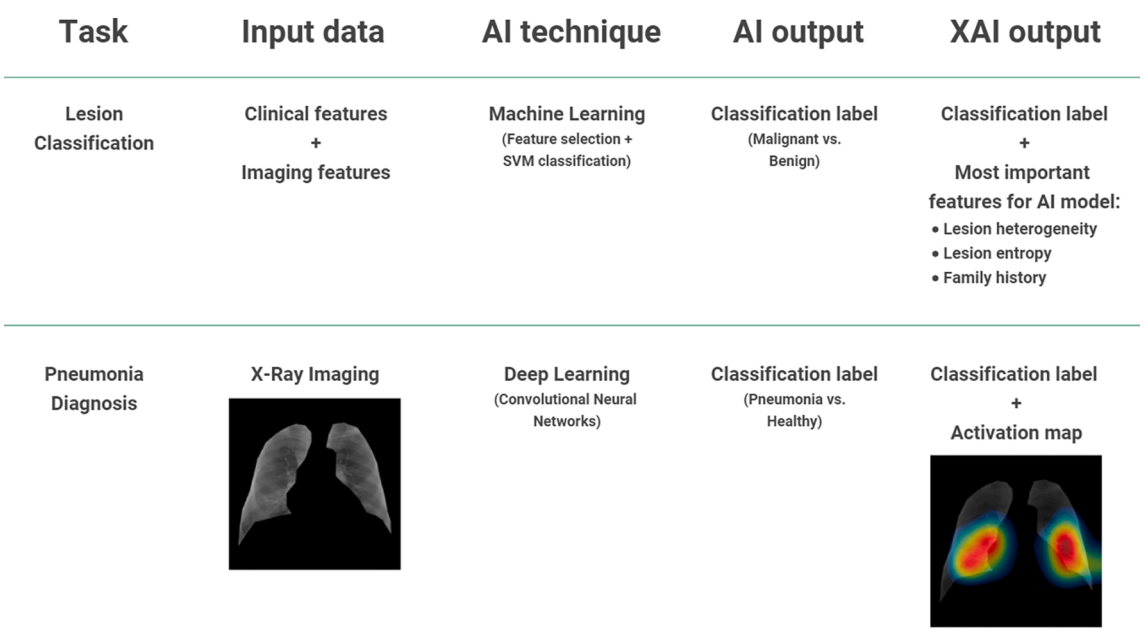


**Fig. 4.** Representative examples of artificial intelligence (AI) tasks in medicine and corresponding AI *versus* explainable artificial intelligence (XAI) outputs.

interpretation of radiomic features and the need for a biological validation of the found radiomics-based biomarkers [156,157]. Given the amount and heterogeneity of available ML and DL algorithms, there is no consensus nor standard strategy to implement XAI yet, although some potential frameworks have recently been proposed [158]. XAI strategies can be grouped based on the learning phase they are applied to and, thus, on the information they reveal (the explainable output). The following phases are considered: *feature reduction* (feature extraction and selection); *learning process* (training and prediction); and the *ensemble of feature reduction and learning process*.

Regarding *feature reduction*, feature extraction and selection techniques are often included within ML systems. Reporting the output of this intermediate phase is a way to make the inner mechanisms of the system more intelligible. When the output consists of a ranked set of extracted/selected features to be used as input for training and prediction, the highest-rank features can be interpreted as the most representative among input data as a function of a given metric. However, this can be independent from the specific predictive task of interest, and thus uninformative. For example, most papers report the principal components extracted from the input dataset [159], representing the features with highest variance in the input dataset, independently from group discrimination. Other feature extraction techniques, such as *Independent Components Analysis* carry similar problems. Conversely, feature extraction techniques such as *Partial Least Squares Analysis* or univariate/multivariate techniques, such as Fisher's discriminant ratio or correlation analysis, can take into account the information about group discrimination. Moreover, it must be noted that some feature extraction techniques do not return a ranked list of extracted features, and thus, different explainability strategies should be adopted, such as those described below. The output of this phase can then be returned as a list of the most representative features of the input dataset (particularly useful if the input dataset is composed of non-image variables) or mapped into the original input space (particularly useful if the input dataset is composed of images). These techniques are easy to implement but their level of explainability is low, being limited to the feature extraction/selection phase, thus not explaining the subsequent training and prediction process.

Regarding the *learning process*, training and prediction represent the core of an ML system. To make this phase interpretable by humans, implemented techniques usually produce a score for each input feature according to its importance in the training-and-classification process. In this case, the resulting feature importance is specific for a given AI classification/prediction task. For example, *random forest* applies an internal optimization technique that minimizes or maximizes a given metric (such as Gini impurity or information gain/entropy), thus returning an importance score based on the contribution of each feature in this optimization process [160]. A similar consideration can be made for *decision trees*. For classifiers based on linear or logistic regression, including *ElasticNet* or LASSO, importance scores are represented by the coefficients found for each input variable during the fitting of the considered distribution. In *support vector machines* (SVMs), the weight assigned by the SVM classifier to each training sample can be back-projected to the original feature space, thus resulting into a score that represents the importance of each feature for SVM classification [161]. However, this last technique can be implemented only when a linear kernel is used [162]. In this second phase the output can be also returned as a simple list of features ranked by importance for classification/prediction or mapped into the original input space, for example highlighted by means of heatmaps. These techniques are characterized by low-to-intermediate implementation difficulty and their level of explainability is limited to the training-and-prediction phase, not explaining the feature-extraction/selection process.

Finally, newer approaches aim at *explaining the AI-system behavior as a whole*, considering both ML and DL techniques. Compared to ML, feature extraction/selection and training/prediction are embedded in a wider process, for example an optimization process. This process can be iterated to optimize a given metric (*e.g.*, classification area under the curve) by varying the number of features given as input to the system. Thus, an importance score can be assigned to each feature according to the corresponding value of the optimized metric. As such, this technique can be used independently of the chosen feature extraction/selection/classification technique, turning AI systems into XAI systems. For example, recursive SVMs can be included in this category, as they use an iterative procedure to assign an importance score to each input feature depending on the entire AI-system performance [163,164].

Since deep architectures encompass the entire learning flow, from feature extraction to classification, DL algorithms can also benefit from XAI, considering the high number of layers in DL architectures, which increase unintelligibility for humans. XAI strategies for DL attempt to unveil how image decomposition works at different depths, and to map this information into saliency/activation maps showing which features of a given image contributed most to the decision. The most popular

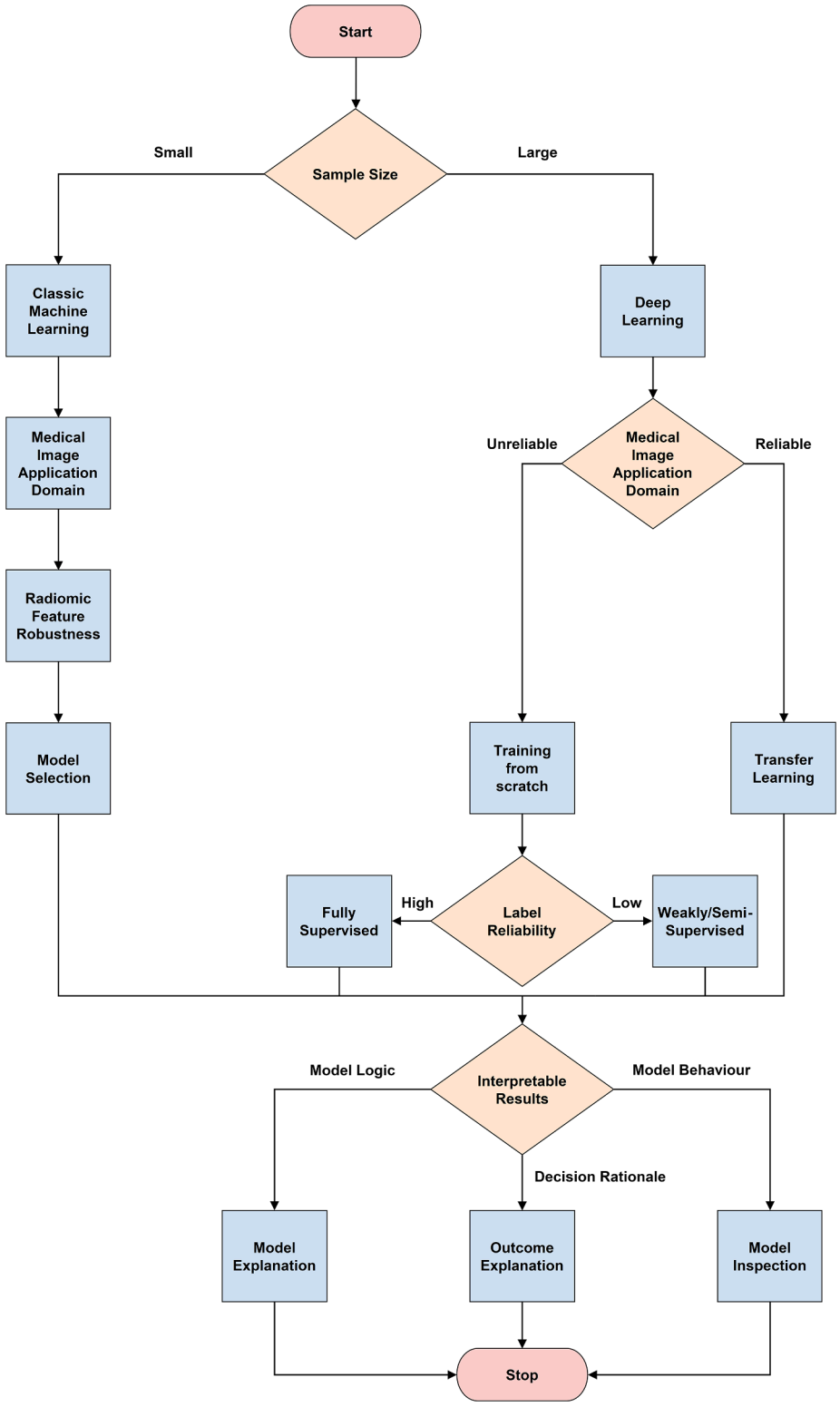


**Fig. 5.** Flow diagram for the design choices in artificial intelligence model development. Each decision block denotes typical practical situations that lead to different solutions in both classic machine learning and deep learning models.

among these techniques is the *Class Activation Map* (CAM) [165–168], in which maps are produced as a function of the pixel-wise activation in the last convolutional layers weighted by the activation contribution to the final score of a given class. CAMs can be generated for any output class, thus returning interpretable information even related to incorrect classification. Another XAI strategy for DL systems has been proposed by Hendricks et al. [169]: the authors trained a CNN to recognize objects in images and a language generating *Recursive Neural Network* was implemented to translate the feature importance of the CNN onto words and captions.

Improving model interpretability represents an open challenge in AI model development to guarantee their translation ability to the clinical

**Table 1**
Challenges of classic machine learning and deep learning models according to decision choices.

| Challenges | Classic Machine Learning | Deep Learning |
|---|---|---|
| Sample size | • Careful radiomic feature robustness and reliability analyses<br>• Strong feature selection process<br>• Machine learning model selection | • *Data augmentation*; *transfer learning*<br>• Regularization to improve model generalizability<br>• *Weakly-*, *semi-*, *self-supervised* or *unsupervised* pre-training<br>• Modify model architecture |
| Medical image application domain | Avoiding dependency on the data *via* careful radiomic feature robustness analyses to avoid *overfitting* on the development set | Use *transfer learning* and domain adaptation to take advantage of pre-trained models or labelled instances from similar domains |
| Label and annotation reliability | • Data curation considering both *segmentation* and response variables<br>• To increase the reliability, multiple labels and morphological perturbations could be considered in the feature robustness analyses | • Data curation considering multicentric and multireader study<br>• Use of image-level labels to derive pixel/voxel-level predictions (*inexact supervision*)<br>• Combine a few well-labelled instances with weakly labelled (*inaccurate supervision*) or unlabeled ones (*incomplete supervision*) |
| Interpretability | High interpretability provided by some models (*e.g.*, *decision trees*) and selected radiomic features (in terms of relevance or importance) | Adopt interpretability and explainability techniques to improve model transparency during both the design and evaluation phases |

domain [170–173]. However, we note that XAI techniques explaining the AI-system behavior as a whole have a high level of implementation difficulty and potentially high computational costs, especially for *wrapped* strategies. Fig. 4 shows two representative examples of possible AI tasks in medicine and the corresponding AI *versus* XAI outputs.

## 6. Design choices: ML *versus* DL

Unfortunately, there is no "*one size fits all*" solution to develop a reliable AI tool. Depending on the quality and quantity of available data, the existence and reliability of labels and *annotations*, as well as the required level of interpretability, AI developers can follow different strategies.

Fig. 5 depicts the main decisions that have to be made during AI model development, involving: (*i*) defining the sample size of the available dataset; (*ii*) assessing whether a previous application domain might be adapted to the problem under consideration, (*iii*) evaluating label and *annotation* reliability; (*iv*) providing interpretability of model results, considering both model logic/behavior and outcome explanation [174]. These choices are better structured in Table 1, which compares classic ML and DL approaches suggesting ways for optimal solutions to the most important challenges of ML and DL applications. However, some commercial and open access software tools already offer multiple functionalities and provide validated solutions for effectively developing AI models in medical imaging without the need of specific AI and coding skills. A classification of these software tools according to their proposed solution to the issues highlighted in this review can be found in the Supplementary Material.

## 7. Closing remarks

In this review, we described the balance between advantages and disadvantages of the use of AI, in particular distinguishing between ML (with its peculiar application to radiomics) and DL. This knowledge is a bridge connecting data scientists (the developers) and clinical users (the physicians) in choosing the best solutions to implement specific AI applications, including special advanced research and immediate clinical needs. Some pros and cons of ML and DL, both specific for each of the two techniques and common to both are presented in Table 2. Four topics deserve a final highlight.

First, when sample size is small, when the predicted class is a label expressed as a continuous variable, or when the integration of additional data (*e.g.*, risk factors or biological data) to imaging features is required

**Table 2**
Pros and cons and recommendations for choosing machine learning or deep learning for application to medical imaging.

| | Pros | Cons | Recommendations* |
|---|---|---|---|
| ML | • A relatively small sample size can be used<br>• Both discrete and continuous variables for *labelling* are possible, eventually with proper feature *oversampling*<br>• Medical image application domain exists and guides the process<br>• (IBSI standardized features for radiomics)<br>• Integration with additional data is possible and easy<br>• High interpretability is immediately provided by some models (*e.g.*, *decision trees*) and is achievable by other algorithms (*e.g.*, SVM) | • Data curation is particularly time-consuming for *image segmentation*<br>• The model must be selected among the possible algorithms (SVM, *random forest*, Bayesian, etc.) | • *Nested* or *wrapped validation* should be performed<br>• Avoiding dependency on the data *via* careful radiomic feature robustness and reliability analyses to avoid *overfitting* on the development set<br>• Apply feature *harmonization*, *intensity normalization*, *denoising*<br>• List the selected features and the most important or relevant features for the model for explainability |
| DL | • Learning curve can be used for stopping sample size<br>• Limited samples can be used but with *transfer learning* or eventually with proper *data augmentation*<br>• Suitable for discrete variables for *labelling*<br>• Medical image application domain exists but does not guide the process<br>• (Use *transfer learning* and domain adaptation to take advantage of pretrained models or labelled instances from similar domains)<br>• *Harmonization*, *intensity normalization*, *denoising* could be avoided if images from variety of datasets are present | • Integration with additional data is possible but very complex<br>• Data curation is particularly time-consuming for *labelling* and *annotations* for image *semantic segmentation*<br>• The ML model must be selected among the possible neural network architectures | • Modify architecture to improve the model performance<br>• Use optimizers in training convergence<br>• Use regularization to improve model generalizability<br>• Provide the saliency map of the activated features for explainability |

ML = machine learning; DL = deep learning; IBSI = Image Biomarker Standardization Initiative; SVM = support vector machines.
*From a general point of view, ensemble learning can be useful in several situations, and the Vapnik–Chervonenkis method can be help sample size definition.

by the model, ML algorithms working in the radiomic domain should be preferred, in agreement with IBSI guidelines. In this case, robust and reliable feature selection, *harmonization* and *denoising*, as well as *nested* or *wrapped validation* schemes, should be performed to avoid *overfitting* and to improve statistical significance of relevant features. Selected relevant features will be the way to explain the model to users.

Second, if a pretrained DL architecture already exists for the specific domain application, transfer learning can be applied as an alternative to radiomics, also in combination with proper *data augmentation*. When this is not possible, if a large and varied sample size is available, DL can be used for training from the scratch. The DL architecture should be modified and adapted to the desired level of feature learning to improve the performance, using optimizers in training convergence and regularization in model generalizability. Saliency maps of the activated features overlapped on original images can explain the model functioning to users.

Third, regarding sample size definition, to avoid subjective assessment and encompass the empirical rule of ten samples per feature, the Vapnik-Chervonenkis method (see section *3.1*) can be used for any AI method, being usefully supplemented by a careful monitoring of the learning curve of the training samples.

Fourth, in addition to performance optimization of individual AI architectures, a better investment by AI developers and users would be the building of combinations of different classifiers, whose overall decision can improve the predictive power of each of them taken individually.

To place the technical and practical knowledge presented in this article into a more general context, we should consider that awareness on the role of AI in human life is only relatively recent. The boost toward a more digital and online world prompted by the COVID-19 pandemic has only exposed a trend in action since 2015, when AI systems started to overcome human readers in image interpretation [175], thanks to the massive increase in computational power we witnessed in the last decade.

Biomedicine and healthcare systems are one of the most important field for AI applications and medical imaging is probably the most suitable and promising domain [35]. When considering the desirable general trend toward the so-called "*P4 medicine*", based on prediction, prevention, personalization and participation, AI tool represents good candidates to facilitate this way to the future [176]. The last "*P*" also, which stands for a more extended patient empowerment, could be increased by a good use of AI, since human intelligence can improve by learning from AI [177], provided that humans have the right knowledge and skills. We can start from healthcare professionals who are facing this unavoidable revolution.

## Conflict of interest

Christian Salvatore is CEO of DeepTrace Technologies S.R.L, a spin-off of Scuola Universitaria Superiore IUSS, Pavia, Italy.

Isabella Castiglioni and Matteo Interlenghi own DeepTrace Technologies S.R.L shares.

Francesco Sardanelli declares to have received grants from or to be member of speakers' bureau/advisory board for Bayer Healthcare, Bracco Group, and General Electric Healthcare.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was partially supported by The Mark Foundation for Cancer Research and Cancer Research UK Cambridge Centre [C9685/A25177]. Additional support has been provided by the National Institute of Health Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2021.02.006.

## References

[1] Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019;380:1347–58. https://doi.org/10.1056/NEJMra1814259.
[2] Rundo L, Militello C, Vitabile S, Russo G, Sala E, Gilardi MC. A Survey on nature-inspired medical image analysis: a step further in biomedical data integration. Fundam Informaticae 2019;171:345–65. https://doi.org/10.3233/FI-2020-1887.
[3] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56. https://doi.org/10.1038/s41591-018-0300-7.
[4] Holzinger A, Plass M, Kickmeier-Rust M, Holzinger K, Crişan GC, Pintea C-M, et al. Interactive machine learning: experimental evidence for the human in the algorithmic loop. Appl Intell 2019;49:2401–14. https://doi.org/10.1007/s10489-018-1361-5.
[5] Rundo L, Pirrone R, Vitabile S, Sala E, Gambino O. Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. J Biomed Inform 2020;108:103479. https://doi.org/10.1016/j.jbi.2020.103479.
[6] Sissons B, Gray WA, Bater A, Morrey D. Using artificial intelligence to bring evidence-based medicine a step closer to making the individual difference. Med Inform Internet Med 2007;32:11–8. https://doi.org/10.1080/14639230601097804.
[7] Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nat Med 2020;26:807–8. https://doi.org/10.1038/s41591-020-0941-1.
[8] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020;26:1364–74. https://doi.org/10.1038/s41591-020-1034-x.
[9] Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26:1351–63. https://doi.org/10.1038/s41591-020-1037-7.
[10] Crigger E, Khoury C. Making policy on augmented intelligence in health care. AMA J Ethics 2019;21:E188–91. https://doi.org/10.1001/amajethics.2019.188.
[11] Kohli M, Alkasab T, Wang K, Heilbrun ME, Flanders AE, Dreyer K, et al. Bending the artificial intelligence curve for radiology: informatics tools from ACR and RSNA. J Am Coll Radiol 2019;16:1464–70. https://doi.org/10.1016/j.jacr.2019.06.009.
[12] Abels E, Pantanowitz L, Aeffner F, Zarella MD, Laak J, Bui MM, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. J Pathol 2019;249:286–94. https://doi.org/10.1002/path.5331.
[13] CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nat Med 2019;25:1467–8. https://doi.org/10.1038/s41591-019-0603-3.
[14] Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, et al. Ethics of artificial intelligence in radiology: summary of the Joint European and North American Multisociety Statement. Radiology 2019;293:436–40. https://doi.org/10.1148/radiol.2019191586.
[15] Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016;278:563–77. https://doi.org/10.1148/radiol.2015151169.
[16] Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006. https://doi.org/10.1038/ncomms5006.
[17] Bishop C. Pattern recognition and machine learning. New York: Springer-Verlag; 2006.
[18] Ranschaert ER, Morozov S, Algra PR, editors. Artificial intelligence in medical imaging. Cham: Springer International Publishing; 2019.
[19] Berenguer R, del Pastor-Juan MR, Canales-Vázquez J, Castro-García M, Villas MV, Mansilla Legorburo F, et al. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. Radiology 2018;288:407–15. https://doi.org/10.1148/radiol.2018172361.
[20] Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EGC, Richter C, et al. Assessing robustness of radiomic features by image perturbation. Sci Rep 2019;9:614. https://doi.org/10.1038/s41598-018-36938-4.
[21] Shafiq-ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E. Voxel size and gray level normalization of CT radiomic features in lung cancer. Sci Rep 2018;8:10545. https://doi.org/10.1038/s41598-018-28895-9.

[22] Scalco E, Belfatto A, Mastropietro A, Rancati T, Avuzzi B, Messina A, et al. T2w-MRI signal normalization affects radiomics features reproducibility. Med Phys 2020;47:1680–91. https://doi.org/10.1002/mp.14038.

[23] Fornacon-Wood I, Mistry H, Ackermann CJ, Blackhall F, McPartlin A, Faivre-Finn C, et al. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. Eur Radiol 2020;30:6241–50. https://doi.org/10.1007/s00330-020-06957-9.

[24] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 2020;295:328–38. https://doi.org/10.1148/radiol.2020191145.

[25] Papanikolaou N, Matos C, Koh DM. How to develop a meaningful radiomic signature for clinical use in oncologic patients. Cancer Imaging 2020;20:33. https://doi.org/10.1186/s40644-020-00311-4.

[26] Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine learning methods for quantitative radiomic biomarkers. Sci Rep 2015;5:13087. https://doi.org/10.1038/srep13087.

[27] Sun P, Wang D, Mok VC, Shi L. Comparison of feature selection methods and machine learning classifiers for radiomics analysis in glioma grading. IEEE Access 2019;7:102010–20. https://doi.org/10.1109/ACCESS.2019.2928975.

[28] Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng 2014;40:16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

[29] Wang L, Ni H, Yang R, Pappu V, Fenn MB, Pardalos PM. Feature selection based on meta-heuristics for biomedicine. Optim Methods Softw 2014;29:703–19. https://doi.org/10.1080/10556788.2013.834900.

[30] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97:273–324. https://doi.org/10.1016/S0004-3702(97)00043-X.

[31] Avanzo M, Wei L, Stancanello J, Vallières M, Rao A, Morin O, et al. Machine and deep learning methods for radiomics. Med Phys 2020;47.. https://doi.org/10.1002/mp.13678.

[32] Castiglioni I, Gallivanone F, Soda P, Avanzo M, Stancanello J, Aiello M, et al. AI-based applications in hybrid imaging: how to build smart and truly multi-parametric decision models for radiomics. Eur J Nucl Med Mol Imaging 2019;46:2673–99. https://doi.org/10.1007/s00259-019-04414-4.

[33] Sala E, Mema E, Himoto Y, Veeraraghavan H, Brenton JD, Snyder A, et al. Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. Clin Radiol 2017;72:3–10. https://doi.org/10.1016/j.crad.2016.09.013.

[34] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749–62. https://doi.org/10.1038/nrclinonc.2017.141.

[35] Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2018;2:35. https://doi.org/10.1186/s41747-018-0061-6.

[36] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[37] Cui S, Tseng H, Pakela J, Ten Haken RK, El Naqa I. Introduction to machine and deep learning for medical physicists. Med Phys 2020;47.. https://doi.org/10.1002/mp.14140.

[38] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.

[39] Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018;9:611–29. https://doi.org/10.1007/s13244-018-0639-9.

[40] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer Vision – ECCV 2014, Cham: Springer International Publishing; 2014, p. 818–33. https://doi.org/10.1007/978-3-319-10590-1_53.

[41] Panayides AS, Amini A, Filipovic ND, Sharma A, Tsaftaris SA, Young A, et al. AI in medical imaging informatics: current challenges and future directions. IEEE J Biomed Heal Informatics 2020;24:1837–57. https://doi.org/10.1109/JBHI.2020.2991043.

[42] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Cham: Springer International Publishing; 2015, p. 234–41. https://doi.org/10.1007/978-3-319-24574-4_28.

[43] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Cham: Springer International Publishing; 2016, p. 424–32. https://doi.org/10.1007/978-3-319-46723-8_49.

[44] Liu L, Cheng J, Quan Q, Wu F-X, Wang Y-P, Wang J. A survey on U-shaped networks in medical image segmentations. Neurocomputing 2020;409:244–58. https://doi.org/10.1016/j.neucom.2020.05.070.

[45] Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 2020;53:5455–516. https://doi.org/10.1007/s10462-020-09825-6.

[46] Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078.

[47] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. arXiv:1406.2661.

[48] Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, et al. GANs for medical image analysis. Artif Intell Med 2020;109:101938. https://doi.org/10.1016/j.artmed.2020.101938.

[49] Ferreira MD, Corrêa DC, Nonato LG, de Mello RF. Designing architectures of convolutional neural networks to solve practical problems. Expert Syst Appl 2018;94:205–17. https://doi.org/10.1016/j.eswa.2017.10.052.

[50] Ruder S. An overview of gradient descent optimization algorithms. arXiv:1609.04747.

[51] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167.

[52] Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. IEEE Access 2019;7:53040–65. https://doi.org/10.1109/ACCESS.2019.2912200.

[53] Moradi R, Berangi R, Minaei B. A survey of regularization strategies for deep models. Artif Intell Rev 2020;53:3947–86. https://doi.org/10.1007/s10462-019-09784-7.

[54] Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Trans Evol Comput 1997;1:67–82. https://doi.org/10.1109/4235.585893.

[55] Jaafra Y, Luc Laurent J, Deruyver A, Saber Naceur M. Reinforcement learning for neural architecture search: a review. Image Vis Comput 2019;89:57–66. https://doi.org/10.1016/j.imavis.2019.06.005.

[56] Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 2020;415:295–316. https://doi.org/10.1016/j.neucom.2020.07.061.

[57] Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. Radiology 2020;295:4–15. https://doi.org/10.1148/radiol.2020192224.

[58] Han C, Murao K, Noguchi T, Kawata Y, Uchiyama F, Rundo L, et al. In: Learning more with less. New York: Association for Computing Machinery; 2019. p. 119–27. https://doi.org/10.1145/3357384.3357890.

[59] Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. In: Kůrková V, Manolopoulos Y, Hammer B, Iliadis L, Maglogiannis I, editors. Artificial Neural Networks and Machine Learning – ICANN 2018, Cham: Springer; 2018, p. 270–9. https://doi.org/10.1007/978-3-030-01424-7_27.

[60] Zhou Z-H. A brief introduction to weakly supervised learning. Natl Sci Rev 2018;5:44–53. https://doi.org/10.1093/nsr/nwx106.

[61] Kervadec H, Dolz J, Tang M, Granger E, Boykov Y, Ben Ayed I. Constrained-CNN losses for weakly supervised segmentation. Med Image Anal 2019;54:88–99. https://doi.org/10.1016/j.media.2019.02.009.

[62] Mao HH. A Survey on Self-supervised Pre-training for sequential transfer learning in neural networks. arXiv:2007.00800.

[63] Ahn E, Kumar A, Fulham M, Feng D, Kim J. Convolutional sparse kernel network for unsupervised medical image analysis. Med Image Anal 2019;56:140–51. https://doi.org/10.1016/j.media.2019.06.005.

[64] Caron M, Bojanowski P, Mairal J, Joulin A. In: unsupervised pre-training of image features on non-curated Data. New York: IEEE; 2019. p. 2959–68. https://doi.org/10.1109/ICCV.2019.00305.

[65] Li Y, Wu F-X, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief Bioinform 2016;19:325–40. https://doi.org/10.1093/bib/bbw113.

[66] Goodfellow I, McDaniel P, Papernot N. Making machine learning robust against adversarial inputs. Commun ACM 2018;61:56–66. https://doi.org/10.1145/3134599.

[67] Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019;363:1287–9. https://doi.org/10.1126/science.aaw4399.

[68] Barucci A, Neri E. Adversarial radiomics: the rising of potential risks in medical imaging from adversarial learning. Eur J Nucl Med Mol Imaging 2020;47:2941–3. https://doi.org/10.1007/s00259-020-04879-8.

[69] Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2:e200029. https://doi.org/10.1148/ryai.2020200029.

[70] Di Leo G, Sardanelli F. Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. Eur Radiol Exp 2020;4:18.. https://doi.org/10.1186/s41747-020-0145-y.

[71] Chan H-P, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers. Med Phys 1999;26:2654–68. https://doi.org/10.1118/1.598805.

[72] Way TW, Sahiner B, Hadjiiski LM, Chan H-P. Effect of finite sample size on feature selection and classification: A simulation study. Med Phys 2010;37:907–20. https://doi.org/10.1118/1.3284974.

[73] Martinetz TM, Ritter HJ, Schulten KJ. Three-dimensional neural net for learning visuomotor coordination of a robot arm. IEEE Trans Neural Networks 1990;1:131–6. https://doi.org/10.1109/72.80212.

[74] Balki I, Amirabadi A, Levman J, Martel AL, Emersic Z, Meden B, et al. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. Can Assoc Radiol J 2019;70:344–53. https://doi.org/10.1016/j.carj.2019.06.002.

[75] Baum EB, Lyuu Y-D. The transition to perfect generalization in perceptrons. Neural Comput 1991;3:386–401. https://doi.org/10.1162/neco.1991.3.3.386.

[76] Fukunaga K, Hayes RR. Effects of sample size in classifier design. IEEE Trans Pattern Anal Mach Intell 1989;11:873–85. https://doi.org/10.1109/34.31448.

[77] Vapnik V, Levin E, Le Cun Y. Measuring the VC-dimension of a learning machine. Neural Comput 1994;6:851–76. https://doi.org/10.1162/neco.1994.6.5.851.

[78] Ojala M, Garriga GC. Permutation tests for studying classifier performance. Journal of Machine Learning Research 2010;11:1833–63.

[79] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019;6:60. https://doi.org/10.1186/s40537-019-0197-0.
[80] Moreno-Barea FJ, Strazzera F, Jerez JM, Urda D, Franco L, Forward noise adjustment scheme for data augmentation.. IEEE Symposium Series on Computational Intelligence (SSCI). New York: IEEE 2018;2018:728–34. https://doi.org/10.1109/SSCI.2018.8628917.
[81] Kang G, Dong X, Zheng L, Yang Y. PatchShuffle regularization. arXiv:1707.07103.
[82] Inoue H. Data augmentation by pairing samples for images classification. arXiv: 1801.02929.
[83] Jia X, Yan H, Cerviño L, Folkerts M, Jiang SB. A GPU tool for efficient, accurate, and realistic simulation of cone beam CT projections. Med Phys 2012;39: 7368–78. https://doi.org/10.1118/1.4766436.
[84] Buvat I, Castiglioni I, Feuardent J, Gilardi M-C. Unified description and validation of Monte Carlo simulators in PET. Phys Med Biol 2005;50:329–46. https://doi.org/10.1088/0031-9155/50/2/011.
[85] Li S, Chen Y, Peng Y, Bai L. Learning more robust features with adversarial training. arXiv:1804.07757.
[86] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. Med Image Anal 2019;58:101552. https://doi.org/10.1016/j.media.2019.101552.
[87] Wolterink JM, Leiner T, Viergever MA, Isgum I. Generative Adversarial Networks for noise reduction in low-dose CT. IEEE Trans Med Imaging 2017;36:2536–45. https://doi.org/10.1109/TMI.2017.2708987.
[88] Calimeri F, Marzullo A, Stamile C, Terracina G. Biomedical data augmentation using Generative Adversarial Neural Networks. In: Lintas A, Rovetta S, Verschure PFMJ, Villa AEP, editors. Artificial Neural Networks and Machine Learning – ICANN 2017, Cham: Springer International Publishing; 2017, p. 626–34. https://doi.org/10.1007/978-3-319-68612-7_71.
[89] Moradi M, Madani A, Karargyris A, Syeda-Mahmood TF. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In: Angelini ED, Landman BA, editors. Proceedings of SPIE Medical Imaging 2018 Image Processing, Bellingham: SPIE; 2018, p. 57. https://doi.org/10.1117/12.2293971.
[90] Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 2018;321:321–31. https://doi.org/10.1016/j.neucom.2018.09.013.
[91] D'Amico NC, Merone M, Sicilia R, Cordelli E, D'Antoni F, Bossi Zanetti I, et al. Tackling imbalance radiomics in acoustic neuroma. International Journal of Data Mining and Bioinformatics 2019;22:365. https://doi.org/10.1504/IJDMB.2019.101396.
[92] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57. https://doi.org/10.1613/jair.953.
[93] Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res 2018;61:863–905. https://doi.org/10.1613/jair.1.11192.
[94] Han H, Wang W-Y, Mao B-H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In: Huang D-S, Zhang X-P, Huang G-B, editors. Advances in Intelligent Computing. ICIC 2005, Berlin, Heidelberg: Springer; 2005, p. 878–87. https://doi.org/10.1007/11538059_91.
[95] Haibo He, Yang Bai, Garcia EA, Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proceedings of the International Joint Conference on Neural Networks, IJCNN 2008, part of the IEEE World Congress on Computational Intelligence, WCCI 2008, New York: IEEE; 2008, p. 1322–8. https://doi.org/10.1109/IJCNN.2008.4633969.
[96] Kotsiantis SB, Pierrakeas CJ, Pintelas PE. In: Preventing student dropout in distance learning using machine learning techniques. Berlin, Heidelberg: Springer; 2003. p. 267–74. https://doi.org/10.1007/978-3-540-45226-3_37.
[97] Liu Xu-Ying, Jianxin Wu, Zhou Zhi-Hua. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics Part B 2009;39:539–50. https://doi.org/10.1109/TSMCB.2008.2007853.
[98] Soda P. An experimental comparison of MES aggregation rules in case of imbalanced datasets. In: 22nd IEEE International Symposium on Computer-Based Medical Systems. New York: IEEE; 2009. p. 1–6. 10.1109/CBMS.2009.5255382.
[99] Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. IEEE Trans Pattern Anal Mach Intell 1998;20:226–39. https://doi.org/10.1109/34.667881.
[100] Soda P, Iannello G. In: A multi-expert system to classify fluorescent intensity in antinuclear autoantibodies testing. New York: IEEE; 2006. p. 219–24. https://doi.org/10.1109/CBMS.2006.21.
[101] Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data, https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf; 2004.
[102] Chen T, Guestrin C. In: XGBoost: a scalable tree boosting system. New York: Association for Computing Machinery; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
[103] Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. Med Phys 2019;46:e1–36. https://doi.org/10.1002/mp.13264.
[104] McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020; 577:89–94. https://doi.org/10.1038/s41586-019-1799-6.
[105] Langlotz CP, Allen B, Erickson BJ, Kalpathy-Cramer J, Bigelow K, Cook TS, et al. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy workshop. Radiology 2019;291: 781–91. https://doi.org/10.1148/radiol.2019190613.
[106] Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Informatics Assoc 2013;20:e147–54. https://doi.org/10.1136/amiajnl-2012-000896.
[107] Imaging 2018;9:1–7.. https://doi.org/10.1007/s13244-017-0588-8.
[108] Spandorfer A, Branch C, Sharma P, Sahbaee P, Schoepf UJ, Ravenel JG, et al. Deep learning to convert unstructured CT pulmonary angiography reports into structured reports. Eur Radiol Exp 2019;3:37. https://doi.org/10.1186/s41747-019-0118-1.
[109] D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA. ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. 5th ed. Reston: American College of Radiology; 2013.
[110] Turkbey B, Rosenkrantz AB, Haider MA, Padhani AR, Villeirs G, Macura KJ, et al. update of prostate imaging reporting and data system version 2. Eur Urol 2019; 2019(76):340–51. https://doi.org/10.1016/j.eururo.2019.02.033.
[111] Irshad H, Montaser-Kouhsari L, Waltz G, Bucur O, Nowak JA, Dong F, et al. Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd. Pacific Symp Biocomput 2015:294–305. https://doi.org/10.1142/9789814644730_0029.
[112] Maier-Hein L, Mersmann S, Kondermann D, Bodenstedt S, Sanchez A, Stock C, et al. Can masses of non-experts train highly accurate image classifiers? In: Golland P, Hata N, Barillot C, Hornegger J, Howe R, editors., Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, Cham: Springer International Publishing; 2014, p. 438–45. https://doi.org/10.1007/978-3-319-10470-6_55.
[113] Kalpathy-Cramer J, Freymann JB, Kirby JS, Kinahan PE, Prior FW. Quantitative imaging network: data sharing and competitive algorithm validation leveraging the Cancer Imaging Archive. Transl Oncol 2014;7:147–52. https://doi.org/10.1593/tlo.13862.
[114] Kalpathy-Cramer J, Beers A, Mamonov A, Ziegler E, Lewis R, Botelho Almeida A, et al. annual meeting. Cancer Imaging Arch 2017;2018. https://doi.org/10.7937/K9/TCIA.2018.OW73VLO2.
[115] Gallivanone F, Interlenghi M, D'Ambrosio D, Trifirò G, Castiglioni I. Parameters influencing PET imaging features: a phantom study with irregular and heterogeneous synthetic lesions. Contrast Media Mol Imaging 2018;2018:1–12. https://doi.org/10.1155/2018/5324517.
[116] Ha S, Choi H, Paeng JC, Cheon GJ. Radiomics in oncological PET/CT: a Methodological Overview. Nucl Med Mol Imaging 2010;2019(53):14–29. https://doi.org/10.1007/s13139-019-00571-4.
[117] Orlhac F, Soussan M, Chouahnia K, Martinod E, Buvat I. 18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. PLoS ONE 2015;10:e0145063. https://doi.org/10.1371/journal.pone.0145063.
[118] Meyer M, Ronald J, Vernuccio F, Nelson RC, Ramirez-Giraldo JC, Solomon J, et al. Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings. Radiology 2019;293:583–91. https://doi.org/10.1148/radiol.2019190928.
[119] Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, et al. Impact of image reconstruction settings on texture features in 18F-FDG PET. J Nucl Med 2015;56: 1667–73. https://doi.org/10.2967/jnumed.115.156927.
[120] Ford J, Dogan N, Young L, Yang F. Quantitative radiomics: impact of pulse sequence parameter selection on MRI-based textural features of the brain. Contrast Media Mol Imaging 2018;2018:1–9. https://doi.org/10.1155/2018/1729071.
[121] Waugh SA, Lerski RA, Bidaut L, Thompson AM. The influence of field strength and different clinical breast MRI protocols on the outcome of texture analysis using foam phantoms. Med Phys 2011;38:5058–66. https://doi.org/10.1118/1.3622605.
[122] Bologna M, Corino V, Mainardi L. Technical Note: Virtual phantom analyses for preprocessing evaluation and detection of a robust feature set for MRI-radiomics of the brain. Med Phys 2019;46:5116–23. https://doi.org/10.1002/mp.13834.
[123] Gibson E, Li W, Sudre C, Fidon L, Shakir DI, Wang G, et al. NiftyNet: a deep-learning platform for medical imaging. Comput Methods Programs Biomed 2018; 158:113–22. https://doi.org/10.1016/j.cmpb.2018.01.025.
[124] Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med 2018;59:1321–8. https://doi.org/10.2967/jnumed.117.199935.
[125] Mahon RN, Ghita M, Hugo GD, Weiss E. ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. Phys Med Biol 2020;65:015010. https://doi.org/10.1088/1361-6560/ab6177.
[126] Messina C, Bignone R, Bruno A, Bruno A, Bruno F, Calandri M, et al. Diffusion-weighted imaging in oncology: an update. Cancers (Basel) 2020;12:1493. https://doi.org/10.3390/cancers12061493.
[127] Dekkers IA, Lamb HJ. Clinical application and technical considerations of T1 & T2 (*) mapping in cardiac, liver, and renal imaging. Br J Radiol 2018;91:20170825. https://doi.org/10.1259/bjr.20170825.
[128] Ma D, Gulani V, Seiberlich N, Liu K, Sunshine JL, Duerk JL, et al. Magnetic resonance fingerprinting. Nature 2013;495:187–92. https://doi.org/10.1038/nature11971.
[129] Truhn D, Schrading S, Haarburger C, Schneider H, Merhof D, Kuhl C. Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast MRI. Radiology 2019;290:290–7. https://doi.org/10.1148/radiol.2018181352.
[130] Xiao D-D, Yan P-F, Wang Y-X, Osman MS, Zhao H-Y. Glioblastoma and primary central nervous system lymphoma: Preoperative differentiation by using MRI-

based 3D texture analysis. Clin Neurol Neurosurg 2018;173:84–90. https://doi.org/10.1016/j.clineuro.2018.08.004.
[131] Huang W, Chen Y, Fedorov A, Li X, Jajamovich GH, Malyarenko DI, et al. The impact of arterial input function determination variations on prostate dynamic contrast-enhanced magnetic resonance imaging pharmacokinetic modeling: a multicenter data analysis challenge. Tomography 2016;2:56–66. https://doi.org/10.18383/j.tom.2015.00184.
[132] Toivonen J, Montoya Perez I, Movahedi P, Merisaari H, Pesola M, Taimen P, et al. Radiomics and machine learning of multisequence multiparametric prostate MRI: Towards improved non-invasive prostate cancer characterization. PLoS ONE 2019;14:e0217702. https://doi.org/10.1371/journal.pone.0217702.
[133] Schwier M, van Griethuysen J, Vangel MG, Pieper S, Peled S, Tempany C, et al. Repeatability of multiparametric prostate MRI radiomics features. Sci Rep 2019;9:9441. https://doi.org/10.1038/s41598-019-45766-z.
[134] Tomasi C, Manduchi R. Bilateral filtering for gray and color images. Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271), Delhi: Narosa Publishing House; 1998, p. 839–46. https://doi.org/10.1109/ICCV.1998.710815.
[135] Buades A, Coll B, Morel J-M. A non-local algorithm for image denoising. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) – 2005;vol. 2:60–5.
[136] Dabov K, Foi A, Katkovnik V, Egiazarian K. BM3D Image denoising with shape-adaptive principal component analysis. In: Gribonval R, editor. SPARS'09 - Signal Processing with Adaptive Sparse Structured Representations. Rennes: Inria; 2009. p. 1–6. https://hal.inria.fr/inria-00369582.
[137] Wu X, Yang Z, Peng J, Zhou J. Global denoising for 3D MRI. Biomed Eng Online 2016;15:54. https://doi.org/10.1186/s12938-016-0168-z.
[138] Tustison NJ, Avants BB, Cook PA, Zheng Yuanjie, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29:1310–20.. https://doi.org/10.1109/TMI.2010.2046908.
[139] Ahlawat S, Fayad LM. Diffusion weighted imaging demystified: the technique and potential clinical applications for soft tissue imaging. Skeletal Radiol 2018;47:313–28. https://doi.org/10.1007/s00256-017-2822-3.
[140] Dietrich O, Biffar A, Baur-Melnyk A, Reiser MF. Technical aspects of MR diffusion imaging of the body. Eur J Radiol 2010;76:314–22. https://doi.org/10.1016/j.ejrad.2010.02.018.
[141] Teoh EJ, McGowan DR, Macpherson RE, Bradley KM, Gleeson FV. Phantom and clinical evaluation of the bayesian penalized likelihood reconstruction algorithm Q.Clear on an LYSO PET/CT system. J Nucl Med 2015;56:1447–52.. https://doi.org/10.2967/jnumed.115.159301.
[142] Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clin Transl Radiat Oncol 2017;4:24–31. https://doi.org/10.1016/j.ctro.2016.12.004.
[143] Pesapane F, Volonté C, Codari M, Sardanelli F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. Insights Imaging 2018;9:745–53. https://doi.org/10.1007/s13244-018-0645-y.
[144] Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2020;2:305–11. https://doi.org/10.1038/s42256-020-0186-1.
[145] Bidgood WD, Horii SC, Prior FW, Van Syckle DE. Understanding and using DICOM, the data interchange standard for biomedical imaging. J Am Med Informatics Assoc 1997;4:199–212. https://doi.org/10.1136/jamia.1997.0040199.
[146] Gambino O, Rundo L, Cannella V, Vitabile S, Pirrone R. A framework for data-driven adaptive GUI generation based on DICOM. J Biomed Inform 2018;88:37–52. https://doi.org/10.1016/j.jbi.2018.10.009.
[147] Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. Npj Digit Med 2020;3:119. https://doi.org/10.1038/s41746-020-00323-1.
[148] Nanayakkara S, Fogarty S, Tremeer M, Ross K, Richards B, Bergmeir C, et al. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. PLOS Med 2018;15:e1002709. https://doi.org/10.1371/journal.pmed.1002709.
[149] Doran D, Schulz S, Besold TR. What does explainable AI really mean? a new conceptualization of perspectives. arXiv:1710.00794.
[150] Adadi A, Berrada M. Peeking Inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138–60. https://doi.org/10.1109/ACCESS.2018.2870052.
[151] Van Lent M, Fisher W, Mancuso M. In: An explainable artificial intelligence system for small-unit tactical behavior. Palo Alto: AAAI Press; 2004. p. 900–7. https://doi.org/10.5555/1597321.1597342.
[152] Andrzejak A, Langner F, Zabala S. Interpretable models from distributed data via merging of decision trees. In: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). New York: IEEE; 2013. p. 1–9. 10.1109/CIDM.2013.6597210.
[153] Piltaver R, Luštrek M, Gams M, Martinčić-Ipšić S. Comprehensibility of classification trees – survey design validation. In: Piltaver R, Gams M, editors. 6th International Conference on Information Technologies and Information Society – ITIS 2014. Ljubljana: Institut Jožef Stefan; 2014. p. 1–16.
[154] Weld DS, Bansal G. The challenge of crafting intelligible intelligence. Commun ACM 2019;62:70–9. https://doi.org/10.1145/3282486.
[155] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Networks Learn Syst 2020. https://doi.org/10.1109/TNNLS.2020.3027314.
[156] Martin-Gonzalez P, Crispin-Ortuzar M, Rundo L, Delgado-Ortet M, Reinius M, Beer L, et al. Integrative radiogenomics for virtual biopsy and treatment monitoring in ovarian cancer. Insights Imaging 2020;11:94. https://doi.org/10.1186/s13244-020-00895-2.
[157] Grossmann P, Stringfield O, El-Hachem N, Bui MM, Rios Velazquez E, Parmar C, et al. Defining the biological basis of radiomic phenotypes in lung cancer. Elife 2017;6:e23421. https://doi.org/10.7554/eLife.23421.
[158] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.
[159] López M, Ramírez J, Górriz JM, Álvarez I, Salas-Gonzalez D, Segovia F, et al. Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease. Neurocomputing 2011;74:1260–71. https://doi.org/10.1016/j.neucom.2010.06.025.
[160] Loh W. Classification and regression trees. WIREs Data Min Knowl Discov 2011;1:14–23. https://doi.org/10.1002/widm.8.
[161] Salvatore C, Cerasa A, Castiglioni I. MRI characterizes the progressive course of AD and predicts conversion to Alzheimer's dementia 24 months before probable diagnosis. Front Aging Neurosci 2018;10.. https://doi.org/10.3389/fnagi.2018.00135.
[162] Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 2014;87:96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067.
[163] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2005;46:389–422. https://doi.org/10.1023/A:1012487302797.
[164] Zhang X, Lu X, Shi Q, Xu X, Leung HE, Harris LN, et al. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. BMC Bioinf 2006;7:197. https://doi.org/10.1186/1471-2105-7-197.
[165] Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Why did you say that? arXiv:1611.07450.
[166] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: 2016 IEEE conference on computer vision and pattern recognition. IEEE; 2016. p. 2921–9. 10.1109/CVPR.2016.319.
[167] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020;128:336–59. https://doi.org/10.1007/s11263-019-01228-7.
[168] Zhao G, Zhou B, Wang K, Jiang R, Xu M. Respond-CAM: Analyzing deep models for 3D imaging data by visualizations. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Cham: Springer; 2018, p. 485–92. https://doi.org/10.1007/978-3-030-00928-1_55.
[169] Hendricks LA, Akata Z, Rohrbach M, Donahue J, Schiele B, Darrell T. Generating visual explanations. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision – ECCV 2016, Cham: Springer; 2016, p. 3–19. https://doi.org/10.1007/978-3-319-46493-0_1.
[170] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15. https://doi.org/10.1038/s42256-019-0048-x.
[171] Baselli G, Codari M, Sardanelli F. Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way? Eur Radiol Exp 2020;4:30. https://doi.org/10.1186/s41747-020-00159-0.
[172] Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nat Commun 2020;11:3673. https://doi.org/10.1038/s41467-020-17478-w.
[173] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Min Knowl Discov 2019;9.. https://doi.org/10.1002/widm.1312.
[174] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2019;51:1–42. https://doi.org/10.1145/3236009.
[175] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115:211–52. https://doi.org/10.1007/s11263-015-0816-y.
[176] Fröhlich H, Balling R, Beerenwinkel N, Kohlbacher O, Kumar S, Lengauer T, et al. From hype to reality: data science enabling personalized medicine. BMC Med 2018;16:150. https://doi.org/10.1186/s12916-018-1122-7.
[177] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500–10. https://doi.org/10.1038/s41568-018-0016-5.