Rundo and Militello *European Radiology Experimental* (2024)8:130
https://doi.org/10.1186/s41747-024-00529-y

EUROPEAN SOCIETY OF RADIOLOGY
European Radiology
EXPERIMENTAL

NARRATIVE REVIEW Open Access

Image biomarkers and explainable AI: handcrafted features *versus* deep learned features

Leonardo Rundo[1]* and Carmelo Militello[2]

Abstract

Feature extraction and selection from medical data are the basis of radiomics and image biomarker discovery for various architectures, including convolutional neural networks (CNNs). We herein describe the typical radiomics steps and the components of a CNN for both deep feature extraction and end-to-end approaches. We discuss the curse of dimensionality, along with dimensionality reduction techniques. Despite the outstanding performance of deep learning (DL) approaches, the use of handcrafted features instead of deep learned features needs to be considered for each specific study. Dataset size is a key factor: large-scale datasets with low sample diversity could lead to overfitting; limited sample sizes can provide unstable models. The dataset must be representative of all the "facets" of the clinical phenomenon/disease investigated. The access to high-performance computational resources from graphics processing units is another key factor, especially for the training phase of deep architectures. The advantages of multi-institutional federated/collaborative learning are described. When large language models are used, high stability is needed to avoid catastrophic forgetting in complex domain-specific tasks. We highlight that non-DL approaches provide model explainability superior to that provided by DL approaches. To implement explainability, the need for explainable AI arises, also through *post hoc* mechanisms.

Relevance statement This work aims to provide the key concepts for processing the imaging features to extract reliable and robust image biomarkers.

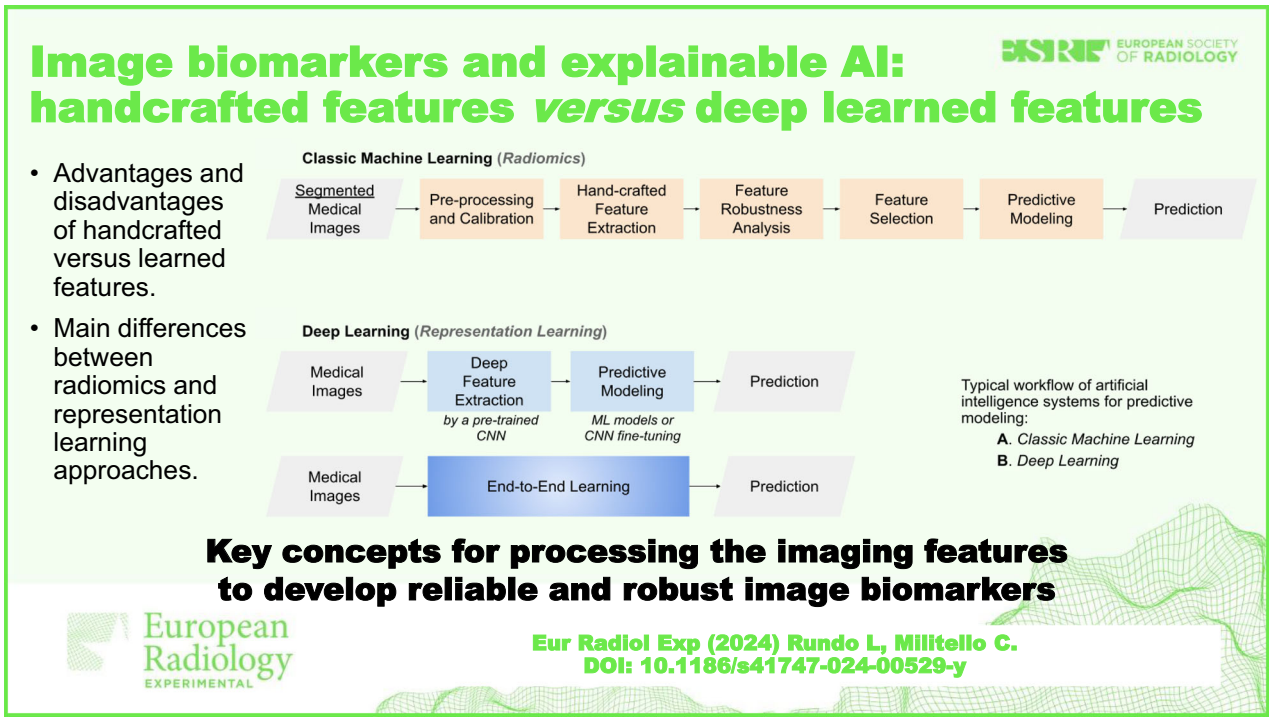
Key Points

- The key concepts for processing the imaging features to extract reliable and robust image biomarkers are provided.
- The main differences between radiomics and representation learning approaches are highlighted.
- The advantages and disadvantages of handcrafted *versus* learned features are given without losing sight of the clinical purpose of artificial intelligence models.

Keywords Biomarkers, Diagnostic imaging, Machine learning, Neural networks (computer), Radiomics

*Correspondence:
Leonardo Rundo
lrundo@unisa.it
[1]Department of Information and Electrical Engineering and Applied Mathematics (DIEM), University of Salerno, Fisciano, Salerno, Italy
[2]High Performance Computing and Networking Institute (ICAR-CNR), Italian National Research Council, Palermo, Italy

Graphical Abstract



Background

Feature extraction and selection, along with the most recent approaches in representation learning, are key steps in the image biomarker discovery process, in many areas of clinical research: oncological imaging [1], cardiovascular imaging [2], and neuroimaging [3]. The purpose of this narrative review is to provide an outline of the main types of methods used in the literature for implementing feature selection. Considering that the literature continuously proposes new methods or improvements of existing methods, providing an exhaustive view would have been impossible, as well as beyond the scope of this paper; it is up to the reader, to investigate the different versions/variants belonging to the specific method. Moreover, an *a priori* assessment of the most suitable feature selection method is difficult, and for this reason, often a 'trial-and-error' approach is exploited [4, 5].

In precision oncology [6, 7], significant advances concerning the definition/identification of new quantitative biomarkers have been obtained through advanced modeling [8] and multimodal data integration [9]. The integration of information coming from different sources makes it possible to improve performance in machine learning (ML). This aspect is particularly true in healthcare applications where clinical and imaging data can be combined. This novel wealth of information that may now be achieved offers unprecedented potential for implementing precision medicine strategies [10] and optimizing the healthcare workflow [11]. However, this type of biomedical image analysis poses unique challenges that must be handled by specific computational approaches [12]. Artificial Intelligence (AI) is emerging as a transformative force in biomedical imaging analysis and has the potential to provide specific support in decision-making processes, enabling strong cooperation between humans and machines, along with performance assessment [13] and clinical decision-making support [14]. Human and machine perceptions are different and sometimes lead to inconsistent results. The case study reported by Makino et al [15] showed that although it is unclear whether humans and deep neural networks (DNNs) use different features to detect microcalcifications, for soft tissue injuries, DNNs rely on high-frequency components ignored by radiologists.

The extraction and computed analysis of imaging-derived features are mandatory phases for proposing accurate and reliable predictive models to be translated and deployed into clinical environments [16]. In this narrative review, we describe: the fundamentals for processing imaging features to extract reliable and robust

image biomarkers; the differences between radiomics based on handcrafted features and representation learning approaches based on learned features; and the advantages and disadvantages of handcrafted *versus* learned features, without losing sight of the clinical purpose of AI models.

Handcrafted features *versus* learned features

Concerning feature extraction and selection, the main issues to consider are the following:

- All extracted features might be dependent on the data source (clinical or imaging data acquisition centers), thus requiring a harmonization step for dealing with distribution shifts in multicenter studies;
- Handcrafted features are strongly dependent on annotations and may not represent the best choice in complex scenarios;
- Learned features, which are directly extracted from deep learning (DL) architectures, generally maximize performance but are affected by the lack of interpretability in clinical practice, thus requiring explanation methods.

When the problem at hand involves an image classification task, the devised classifier has to take as input some data of interest and provide the output. In the setup phase to define the classifier input, there are two options [17, 18]: (1) using the raw data, as the original pixel/voxel values; (2) extracting features from the image employing well-defined mathematical formulations. Choosing the second approach (extracted features), it is possible to:

- Arbitrarily and manually define a set of features, even though we do not know *a priori* if the selected features are appropriate for the specific classification task (this approach has been used for many years by using classic ML techniques after the feature extraction step);
- Train an ML model to extract and identify useful features for the specific classification task: this approach has become predominant with the advent of DL, which can learn the optimal set of features [19].

More precisely, since the learned features are extracted automatically to solve a specific task, they are extremely effective at it, typically allowing DL models to outperform classic ML models based on handcrafted features manually extracted by the data scientist or AI engineer [20]. This trend is most often observed when comparing ML and DL, but we note that, depending on the scenario and data, classical ML allows for comparable performance to DL architectures.

Regardless of the approach used (ML or DL), the type and distribution variability of input/output data can affect the operation and, consequently, the model performance. ML techniques used in computer-aided medical image analysis usually suffer from the domain shift problem, caused by the different distributions between the source/reference data and the target data. This aspect, which is very evident in multicenter studies, must be considered. To this end, domain adaptation has attracted considerable attention in recent years [21].

While there is the advantage of obtaining features learned automatically by the classifier and being able to develop effective models, there is no control over which features the model will extract from the data and their meaning. Deep features, another way learned features are called because they are intrinsically related to the convolutional layers of convolutional neural network (CNN) architectures [22], are effective for the task under consideration. However, they do not always provide a direct real-world interpretation while the scientific and medical community is devoting its attention to explainable approaches [23].

It is generally difficult to identify the most suitable approach (handcrafted *versus* deep features), and researchers often experimentally compare both alternatives [17].

For the sake of completeness, we highlight that ML algorithms can be categorized into two fundamental types: (1) supervised learning; and (2) unsupervised learning. In supervised learning, there is always a specific measure to be predicted: the target, which represents the dependent variable. All other variables—the ones used to predict what the target will be—are called features and are the independent variables given as input to the model. Depending on the nature of the target variable, you can identify two scenarios of learning that require different learning algorithms within the supervised family: (1) when the target variable is a numeric measure, you need a regression algorithm; (2) when the target variable is a categorical measure, you will need a classification algorithm. In unsupervised learning, the objective is not to make a prediction but to unveil some hidden structure in your data. Unsupervised ML algorithms are capable of exploring your data to find some interesting patterns, a finality that is often called clustering [24].

Radiomics

Leveraging medical imaging, radiomics is a technique allowing the extraction of features [25] that can be mined to noninvasively assess the *in vivo* phenotype of lesions or even just certain tissue parts (*e.g.*, the normal tissue surrounding a tumor) [26–28]. The segmentation of the regions of interest (ROIs) or volumes of interest (VOIs) containing the area/volume from which to extract radiomic features—a process that can be performed through

Table 1 Classes of handcrafted features along with their types and descriptions

Class of handcrafted features	Types and details
First-order	Histogram-derived features [93]: describe the distribution of voxel intensities within the image ROI
Second-order	Gray-level co-occurrence matrix (GLCM) features: quantify the spatial relationship between pixels, revealing homogeneity, uniformity, linear dependencies, and randomness [94, 95]
Higher-order	Gray-level run-length matrix (GLRLM) features: quantifies gray-level runs, which are defined as the length in number of pixels, of consecutive pixels that have the same gray-level value [67, 96, 97]
	Neighboring gray-level dependence matrix (NGLDM): quantify pixel properties invariant under rotation; and linear gray-level transformation, useful for texture description and classification [98]
	Neighborhood gray-tone difference matrix (NGTDM): quantifies the difference between a gray value and the average gray value of its neighbors within a specific distance, able to approach human perception [99]
	Gray-level size zone matrix (GLSZM): quantifies gray-level variations and homogeneity of image zones
	Gray-level distance zone matrix (GLDZM): assesses zones of neighboring pixels or voxels with the same gray level and at the same distance from the ROI boundary [100]
	Gray-level dependence matrix (GLDM): quantification of gray-level dependencies [98]
Transformed domain	Absolute gradient matrix (AGM): texture extraction and analysis
	Histograms of oriented gradients (HOG): considering the occurrences of gradient orientation in localized portions of an image [101]
	Local binary patterns (LBP): assign binary numbers to each pixel in an image by thresholding its surrounding pixels [102]
	Gabor transform [103]
	Wavelet transform [104]

manual or computer-assisted (automatic/semiautomatic) procedures. The predictive models using traditional ML techniques begin with the extraction of large-scale handmade radiomic characteristics: morphometric features (*e.g.*, size, shape, diameter measurements) and function quantifying tissue textures (*e.g.*, first-order, second-order, higher-order, and transformed domain descriptors). Starting from an input ROI/VOI, the radiomic features can be calculated in two ways: (1) voxel-based extraction (for each feature, a value is computed for each voxel, thus yielding feature maps as output); (2) segment-based extraction (a single, aggregated value per feature is computed for each ROI/VOI). The main classes of handcrafted features are listed in Table 1.

Although radiomic features are well-known among handcrafted features in the Computer Vision community [25, 29], there are still serious concerns about their stability and robustness [30, 31]. Indeed, radiomic features suffer from a lack of robustness against imaging parameters (*e.g.*, spatial resolution) [32] and image extraction settings (*e.g.*, quantization levels, resampling) [33–35]. Furthermore, also the software used to extract radiomic characteristics may have an impact on them [36].

The Image Biomarker Standardisation Initiative (IBSI) [29], which provides definitions and nomenclature of radiomics characteristics and defines computation and normalization procedures, tries to resolve/alleviate outstanding challenges in this area. The IBSI also offers implementation recommendations for the various steps of a radiomic workflow, such as intensity discretization, re-segmentation, postacquisition image processing, segmentation, data interpolation, and data conversion in standardized units [37]. Furthermore, for every feature in a model based on binary classifiers, at least ten samples (*i.e.*, patients) would be required, according to a well-known rule of thumb [25]. Consequently, adding more features could exacerbate the curse of dimensionality issue by adding redundancy to the features, particularly when there are not enough samples available.

To establish robust imaging biomarkers, feature selection procedures specific to the radiomics domain must be developed once features have been calculated and normalized [1]. To achieve this, the selection process should remove: (1) unreliable features (using the intraclass correlation coefficient, for example); (2) features that are not informative based on zero or nearly zero variance; and (3) redundant features (such as those that have a high correlation with one another). Figure 1 outlines the typical preprocessing and calibration steps required by a robust radiomics pipeline as outlined in the following paragraphs.

Feature robustness analysis

It is aimed at identifying robust features, for example, considering the effect of variability of the ROIs, *i.e.*, the natural situation due to intra- and inter-reader dependence during manual contouring [38]. Moreover, also the optimal

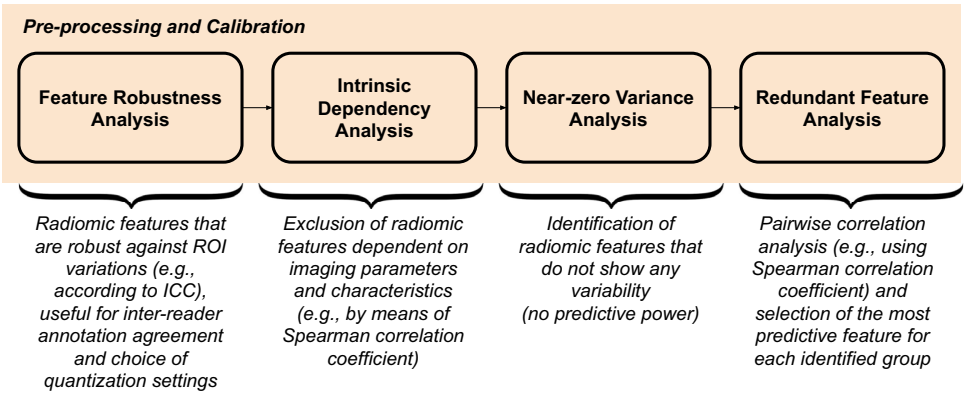


Fig. 1 Preprocessing and calibration of radiomic features is a mandatory step to obtain a subset of features that are independent of the region/volume of interest segmentation and the imaging acquisition and reconstruction parameters, which are relevant from the point of view of information content, and nonredundant features. ICC, Intraclass correlation coefficient

quantization setting can be evaluated by extracting the radiomic features considering different quantization levels in terms of either number of bins or bin width. According to the IBSI guidelines, the former (the number of bins) is recommended for calibrated imaging ranges (*e.g.*, Hounsfield units in computed tomography, standardized uptake value in positron emission tomography), while the latter (the bin width) is generally used in non-standardized ranges (*e.g.*, images obtained through common non-quantitative magnetic resonance imaging sequences). The intraclass correlation coefficient analysis can be used to take into account the extracted features and allows us to determine which are more robust as the perturbations vary and the number of bins varies [39].

Intrinsic dependency analysis

Imaging characteristics, and consequently the extracted features, can vary as a function, for example, of magnetic resonance imaging acquisition parameters, such as scanner type, scanner setting, and imaging protocols [40]. For this reason, a correlation analysis (*e.g.*, using the Spearman correlation coefficient) allows us to evaluate which features are correlated with, and perhaps dependent on, the imaging acquisition parameters and to select only those features that do not appear to be dependent on them.

Near-zero variance analysis

Its goal is to eliminate characteristics that do not convey high information content. This procedure takes into account a cutoff for the percentage of distinct values across all samples as well as a cutoff for the ratio of the most frequent value to the second most common value.

Redundant feature analysis

It is aimed at removing highly correlated features to reduce the redundancy among the features. In the case of a correlation value (*e.g.*, Spearman correlation coefficient) higher than a threshold (0.90 is a widely used value), the feature with the highest predictive power is selected. A simple skimming approach can be performed by a univariate logistic regression for predicting the lesion characterization, by removing the feature that achieved the lowest area under the receiver operating characteristic. A more sophisticated approach might rely upon dendrograms for identifying groups (*i.e.*, clusters) of highly correlated features.

More recently, the dependence on acquisition and reconstruction parameters, such as different reconstruction kernels in computed tomography, has been very recently addressed by using generative adversarial networks (GANs) for image-to-image translation [41, 42]. By doing so, the heterogeneity of the dataset can be mitigated, or even missing data can be generated across sequences or modalities [43, 44]. GAN-based image synthesis has been also combined with diffusion models [45], which represent the latest developments for unsupervised generative approaches in medical imaging [46]. Data harmonization, which allows for the adjustment of variations in imaging methods that generally produce noise in non-AI imaging research, is also crucial in the context of multicentric and multi-institutional studies. With more details, these techniques maintain the information content of images by normalizing the statistical distributions of the same attributes when they are acquired from other systems, such as in the case of ComBat [47], able to combat the batch effect, and its generalizations [48].

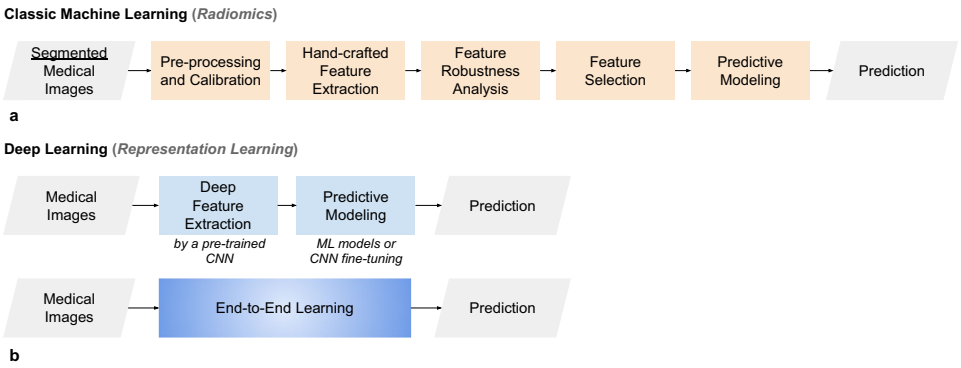


Fig. 2 Typical processing pipelines of AI systems for the implementation of predictive models. **a** Classic ML, including the various steps that process handcrafted features. **b** DL exploits representation learning by relying upon deep image feature extraction or end-to-end learning. AI, Artificial intelligence; CNN, Convolutional neural network; DL, Deep learning; ML, Machine learning

Attempts in pictorial interpretation [49], as well as biological and validation, of the radiomic signatures [50], have been carried out. The main characteristic is the intelligibility of radiomic features. Indeed, shallow learning and explainable methods provide insights into the features driving their decisions, allowing clinicians to validate the reasoning behind the recommendation of the system.

Finally, radiomic feature extraction has shown several advantages over deep feature extraction. It is possible to perform an accurate feature extraction also on moderate sample sizes, while deep feature extraction requires a large database to avoid the overfitting issue [51].

Deep learning models

With DL models, it is possible to optimize the model's performance for the given task by automatically extracting image features. DL is a particular sub-field of ML that uses artificial neural networks to interpret raw data directly [52]. To provide the deep architecture with appropriate input, some preprocessing steps must be implemented to condition the data appropriately, and to make the images suitable for deep learning models to process (*e.g.*, intensity scaling, resizing, patching, etc). All these processes can impact the performance of the DL model and, for this reason, should be carefully dimensioned and defined case-by-case, depending on the scenario's data. DNNs make it possible to create end-to-end prediction models by handling every processing step—including feature extraction and learning—that is often required to create a traditional ML model (Fig. 2).

DNNs provide representation learning techniques based on a stack of processing layers with a finite number of nonlinear units (*i.e.*, artificial neurons) [53, 54]. Input and output layers are the top and bottom levels of the network, respectively, and the layers in the middle are called hidden layers [55]. DNNs can act as nonlinear function approximators because of their multilayered structure, which allows them to learn many representations of the input data at different levels of abstraction [54]. In contrast, a single neuron acts as a linear classifier.

A DL model can quickly reach millions of trainable parameters to estimate during the training phase, depending on the number of layers and units per layer. As a result, DL models need datasets containing thousands of images, as they are prone to overfitting, especially when working with relatively small training sets [56]. Applications specific to the field of medical imaging encompass both small and big imaging databases, albeit with different consequences. Since DL can predict incredibly complex relationships within large datasets, it has found widespread application in radiation oncology and medical imaging [57].

Although ROI/VOI segmentation is not needed for extracting these features (depending on the task at hand), the effectiveness of CNN-based classification might be affected by the different image or patch sizes for including the whole ROI. Importantly, CNNs allow for a hierarchical abstraction of the input data, thus resulting in highly effective signals characterized by a strong spatial/temporal continuity, such as in the case of multidimensional images: CNNs work hierarchically so that the output of one convolution layer is used as input in the next convolution layer. CNN is a type of artificial neural network in which the pattern of connectivity between neurons is inspired by the organization of the human visual cortex, whose individual neurons are arranged in such a way as to respond to the overlapping regions that tessellate the visual field [55]. A typical CNN-based architecture is represented in Fig. 3. According to the architecture shown in Fig. 3, the main components of CNNs are reported in Table 2.

Tuning the hyperparameters is a nontrivial step after choosing the best network architecture. Structural

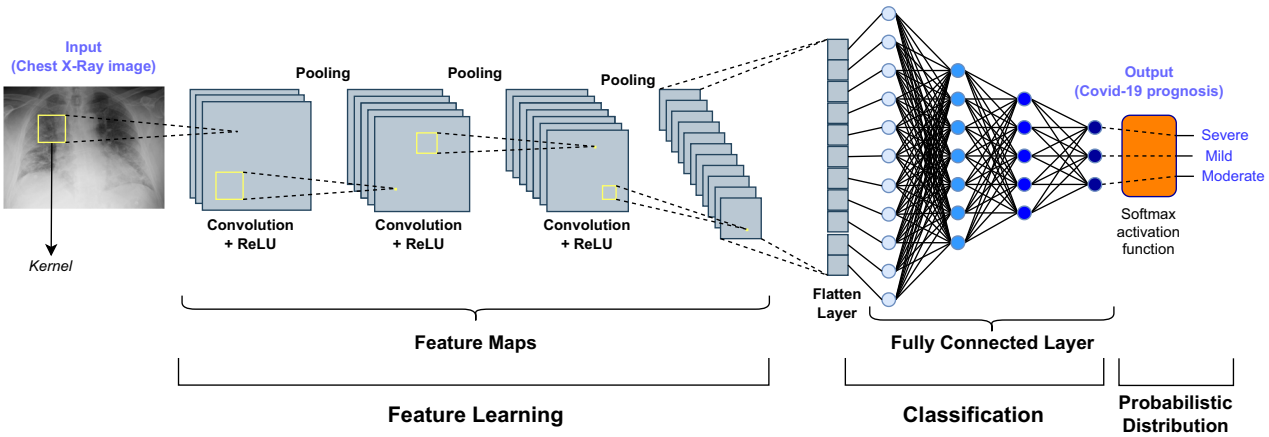


Fig. 3 Typical architecture of a CNN for an image classification task. The layers of a CNN learn the features and the final classification is performed *via* a fully connected layer, which processes the feature maps after a flattening operation. In this case, an example of COVID-19 prognosis (*i.e.*, three-class classification) based on the analysis of chest x-ray images is depicted [105]. CNN, Convolutional neural network; ReLU, Rectified linear unit

hyperparameters that can greatly affect model performance include the number of layers/neuronal units, the activation functions, and the receptive field size (the area of the input space that a particular CNN's feature is interested in). This makes designing the optimal architecture challenging [53].

Representation learning

Representation learning was introduced to represent the information efficiently and learn the abstract features from the raw data [58]. These techniques, based on DNNs, are different from handcrafted feature extraction and selection and allow the raw input data to be converted into meaningful intermediate features and outputs [59].

Transfer learning—an ML technique in which a pre-trained model on a task is tuned to a new, related task—can play a crucial role in feature learning, by allowing also for domain adaptation across different institutions [60]. The simplest approach for transfer learning involves "freezing" the first hidden layers (*i.e.*, the weights will not be trained, but will keep the values from the previous training procedure) and then applying fine-tuning [61]. Interestingly, shared factors across tasks may also exist, such as in the case of joint segmentation and classification tasks. With many inter-related tasks of interest or many learning tasks in general, each task can be explained by factors that are shared with other tasks, thus allowing for the sharing of statistical strengths across inter-related tasks [58].

CNNs have also been used in conjunction with recurrent neural networks to extract temporal and spatial information from imaging data series. These networks allow the processing of new data (such as longitudinal image series of arbitrary length) while keeping track of previous inputs and outputs since they share node weights throughout time. However, because model complexity is directly correlated with the amount of input data, recurrent neural networks are challenging to train and prone to overfitting.

Feature selection and dimensionality reduction

In ML, we often incur the so-called problem of the curse of dimensionality, where the number of data samples (*e.g.*, patients) is substantially lower than the number of features processed. This could cause problems because it calls for training a large number of parameters on a sparse dataset, which increases the risk of overfitting and suboptimal generalization. However, high dimensionality also leads to extremely long training durations. Therefore, to solve these problems, dimensionality reduction techniques are often used in addition to feature selection. Moreover, while residing in an initial high-dimensionality space, the final space of features has a lower-dimensional structure.

Dimensionality reduction is the transformation from a high-dimensional space (*i.e.*, the dataset space) into a low-dimensional space so that the low-dimensional representation retains only meaningful properties of the original dataset. In radiomic-based approaches, after the preprocessing and calibration steps described in Fig. 1, a subsequent selection step is carried out aimed at identifying the most relevant predictive features [62, 63]. In quantitative imaging and radiomics, this approach is preferred to dimensionality reduction since the input features are preserved and are not involved in any transformation process [64]. This aspect is crucial in the development of interpretable models that rely upon the meaning of handcrafted radiomic features.

The high dimensionality of datasets in radiomic investigations is a significant issue, resulting from limited

Table 2 Main components of a CNN architecture, describing functioning details and processing output

CNN component	Details	Output
Convolutional layers	Kernel size, striding, padding, receptive field. Smaller kernels (3 × 3) with many convolutional layers (~ 1 K) imply fewer parameters than fully connected nets.	Convolutional layers extract features. In general, more than one kernel is applied, thus obtaining a different feature map. Given a square image as input composed of $W_{in} \times W_{in}$ pixels, N different kernels with dimension $K \times K$, stride S, and padding P, the output will have dimensions $W_{out} \times W_{out} \times N$, where $W_{out} = [(W_{in} - K + 2P)/S] + 1$.
Pooling layers	Translation-invariant (preserving important information in local patches). • Max pool (preferable in middle layers) • Average pool (most used in the final layers)	Pooling layers perform a downsampling by dividing the input into regions (*i.e.*, pooling windows) and performing an aggregation operation, such as taking the maximum or average value, within each window. This aggregation reduces the size of the feature maps, resulting in a compressed representation of the input data.
Dropout	Destroy/preserve connections with output neurons randomly (with probability $p_{dropout}$, which is a hyperparameter used also at test time). • This does not increase the number of parameters and acts as a regularizer • Equivalent to training several smaller networks (in terms of average features from smaller nets) • Training time increases, but it is reduced compared to a multi-expert approach	Dropout performs a regularization for reducing overfitting and improving the generalization of CNNs. Connections are dropped out with a rate $p_{dropout}$ at each step during training time.
Activation functions	Sigmoid, softmax, linear, rectified linear unit (ReLU), hyperbolic tangent (Tanh), leaky ReLU	Activation function is used to determine the output of neural network (*e.g.*, yes/no, class A/class B, …). The function maps the resulting values in between 0 and 1 or -1 and 1, etc. (depending the used function).
Output layers	Typically fully connected (like in traditional multilayer perceptrons), also called dense layers • Binary classification: one node, sigmoid activation • Multiclass classification: one node per class, softmax activation • Multilabel classification: one node per class, sigmoid activation • Regression: one node, linear activation	Output layers are typically fully connected (FC) layer, called that because each neuron from the previous layer is connected to each neuron of the current layer. FC layers—typically found towards the end of a neural network architecture—are responsible for producing final output predictions.

sample numbers and a large number of generic features retrieved from the VOI. To eliminate unnecessary and redundant features, feature selection techniques are applied. Given the abundance of feature selection algorithms available, it is critical to comprehend each one's effectiveness in the context of radiomics [65].

All of these methods can address the "curse of dimensionality" and decrease overfitting in the model, improving the model's capacity for generalization. Three classes of feature selection techniques exist:

(i) Filter methods, which evaluate a feature subset's usefulness using information theory-based metrics or statistical correlation;
(ii) Wrapper methods, which use a search method (such as recursive feature elimination, sequential feature selection, or metaheuristics) to evaluate feature combinations and maximize the predictive model's performance;
(iii) Embedded methods, which enable feature selection during the model's training, as in the cases of Elastic net regularization techniques (ElasticNet) and Least Absolute Shrinkage and Selection Operator (LASSO).

Among these, wrapper methods are powerful but computationally demanding [66]. To find the best feature subset, they rely on the evaluation of classification performance. Since exhaustive search methods are computationally intensive and impractical for large-scale datasets, search methods and metaheuristics are typically

used to find suboptimal solutions in the search space [67]. Most importantly, repetitive statistical comparisons may create overfitting in the feature subset space as a result of the repeated accuracy estimation utilized in feature subset selection, which would hinder generalization skills [68]. The quality, diversity, and quantity of the data used can directly impact the reliability of the results obtained and can limit the model's generalization ability. Indeed, medical imaging tasks are typically affected by noise, missing data, and class imbalance since pathological samples represent the minority class compared to healthy samples [69]. Therefore, resampling methods are fundamental for handling missing data during the data curation phase, as well as for dealing with highly unbalanced data during the ML-based modeling phase (*i.e.*, data augmentation *via* minority class oversampling) [70].

The output of the selection processes is a set containing relevant, nonredundant, and robust features. The next step of the pipeline is the definition of the predictive model. Depending on the specific clinical topic at hand, multivariable classification or regression techniques can be used to do this [71], usually in supervised learning environments. Interestingly, unsupervised feature selection methods [72] are effective and robust in radiomics applications [73].

The training durations, stability, and similarity of feature selection techniques varied significantly [65]: no single prediction technique was able to consistently outperform the others. According to these findings, less complicated techniques outperform more complicated ones in terms of the area under the receiver operating characteristic curve [74]. They are also more stable. In terms of predictive performance, analysis of variance, LASSO, minimum redundancy, and maximum relevance ensemble seem well-suited for radiomic research, as they outperformed the majority of other feature selection techniques.

Principal component analysis

To acquire lower-dimensional data while retaining as much of the variation in the data as feasible, the principal component analysis is frequently employed for dimensionality reduction [75]. This is achieved by projecting each data point onto the first principle components only. A direction that optimizes the projected data's variance can be used to define the first main component. Generally speaking, the direction orthogonal to the first (i–1) main components that maximizes the variance of the projected data is the principal (i-th) component.

To summarize, principal component analysis learns a linear transformation by projecting the input data onto another space. By limiting dimensionality to some components according to the explained variance of the dataset, dimensionality reduction can be achieved.

t-distributed stochastic neighbor embedding (t-SNE)

The t-SNE is a nonlinear dimensionality reduction technique designed to effectively embed high-dimensional data for visualization in a two- or three-dimensional low-dimensional environment [76, 77]. To be more specific, every high-dimensional object is represented by t-SNE as a two- or three-dimensional point, with a high probability of representing related objects by nearby points and dissimilar objects by distant points. There are two primary steps in the t-SNE algorithm: (1) creation of a probability distribution between pairs of high-dimensional objects in which the likelihood of similar objects is higher and the probability of different points is lower; (2) establishment of an equivalent probability distribution on the low-dimensional map points and reduction of the Kullback-Leibler divergence between the two distributions concerning the map points' locations.

The original algorithm bases its similarity metric on the Euclidean distance between objects, although this can be changed as needed.

Uniform manifold approximation and projection (UMAP)

Similar to t-SNE, the UMAP technique [78] reduces nonlinear dimensionality and can also be applied to generic nonlinear dimension reduction. The three basic assumptions of the UMAP algorithm (concerning data within the Riemannian geometry) are (1) uniform distribution of the data; (2) locally constant Riemannian metric (or approximable); and (iii) locally connected manifold.

Autoencoders

Autoencoders are an important kind of DL architecture that may be used to reduce the input into a low-dimensional latent space [79, 80]. These networks use progressively smaller hidden layers in the encoder path, regularization, and sparsity constraints to enable learning a lower-dimensional representation of the data, preventing the network from learning the identity transformation, which copies the source data into the destination data without alteration (*i.e.*, the trivial solution) [57].

By stacking several nonlinear transformations, each autoencoder layer handles a different transformation. Their design consists of an encoder-decoder, in which the input is mapped to latent space by the encoder and then reconstructed by the decoder. The back-propagation process is used to train them so they can correctly reassemble the input. Autoencoders can be used for dimensionality reduction when the latent space has smaller dimensions than the input. It makes sense that the most significant characteristics of the particular application are encoded by these low-dimensional latent variables, which are discovered during the reconstruction process.

Conclusions

We summarized and discussed the main aspects of feature extraction and selection with a particular interest in the extraction of reliable and robust biomarkers. We have shown that there is no unifying technique yet. Therefore, despite the outstanding performance of DL methods in many medical image analysis tasks, the use of either handcrafted or learned features needs to be carefully considered for each different study. Radiomics-powered analyses still play a key role in clinically feasible and interpretable applications, allowing for studies that rely on datasets with a limited number of cases.

The size of the dataset is a key aspect: obtaining datasets with too many or too few cases does not represent an optimal situation for model setup. In fact, large-scale datasets (with low sample diversity) could lead to model overfitting; conversely, datasets with limited sample sizes can provide unstable models. Each application scenario must be evaluated in-depth to define the amount of data needed to obtain a well-trained and reliable model. Undoubtedly, the dataset must be representative of all the 'facets' of the clinical phenomenon (*i.e.*, disease) under investigation.

The access to the computational resources (*i.e.*, hardware and software) needed for methods requiring high computational performance hardware—which only graphics processing units can provide—could be a limitation for rapid and broad implementation of the studies (and for the proposed methodologies). This problem is mostly evident in the training phase (once trained, computational demands decrease significantly) and for solutions that rely on deep architectures. However, it must be pointed out that the market now offers very high-performance graphics processing units at affordable prices and programmable with different software, many of them open-source.

Data engineering of large-scale datasets will be fundamental to developing accurate, generalizable DL-powered methods in the near future. A new perspective to avoid data sharing and privacy protection is the federated learning paradigm, also known as collaborative learning, an ML technique allowing an algorithm to be trained through the use of decentralized devices or servers that store data [81, 82]. It enables multi-institutional and reliable studies, along with appropriate data harmonization techniques for information fusion [83]. This aspect has to be taken into account. In fact, it could be very useful in healthcare applications where, because of the sensitive data used, data managers (*e.g.*, hospitals) put constraints on data transfer [84]. Moreover, features that are peculiar and well-established for image biomarkers might be effectively supported by large language models (LLMs) [85] in the case of interactive diagnostic tasks powered by AI tools [86, 87].

Indeed, the introduction of LLMs could effectively support diagnostic tasks in quantitative imaging, such as radiology reporting [88], after careful evaluation [89]. LLMs exhibit great performance in language understanding and generation. However, when LLMs are fine-tuned on complex domain-specific tasks, their inference performance on past/historical tasks decreases dramatically [90]. This drawback—called catastrophic forgetting—refers to a phenomenon where an LLM tends to lose previously acquired knowledge as it learns new information. This aspect represents a 'drift' for the model and must be addressed to have LLM with stable performance [91, 92].

In conclusion, shallow-learning approaches can provide model explainability that is difficult to achieve with deep architectures. This criticality stems from two different aspects: (1) the lack of interpretability of learned features and (2) the cryptic operation mechanism of deep architectures. So, in DL models, in the face of better performance, we go to a loss of explainability. From this issue, the need for explainable AI arises, which aims to implement explainability (*e.g.*, through *post hoc* mechanism) within ML models.

Abbreviations

AI	Artificial intelligence
CNN	Convolutional neural network
COVID-19	Coronavirus disease 2019
DL	Deep learning
DNN	Deep neural network
IBSI	Image Biomarker Standardization Initiative
LLM	Large language model
ML	Machine learning
ROI	Region of interest
t-SNE	t-distributed stochastic neighbor embedding
UMAP	Uniform manifold approximation and projection
VOI	Volume of interest

Acknowledgements

No AI-assisted technologies for writing, such as Large Language Models, were used for this manuscript.

Author contributions

Both authors—LR and CM—contributed equally to all phases of the study, from design to writing, to revision and final proofreading.

Funding

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

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 6 April 2024 Accepted: 16 October 2024
Published online: 19 November 2024

References

1. Papanikolaou N, Matos C, Koh DM (2020) How to develop a meaningful radiomic signature for clinical use in oncologic patients. Cancer Imaging 20:33. https://doi.org/10.1186/s40644-020-00311-4
2. Guglielmo M, Lin A, Dey D et al (2021) Epicardial fat and coronary artery disease: role of cardiac imaging. Atherosclerosis 321:30–38. https://doi.org/10.1016/j.atherosclerosis.2021.02.008
3. Young PNE, Estarellas M, Coomans E et al (2020) Imaging biomarkers in neurodegeneration: current and future practices. Alzheimers Res Ther 12:49. https://doi.org/10.1186/s13195-020-00612-7
4. Demircioğlu A (2022) Evaluation of the dependence of radiomic features on the machine learning model. Insights Imaging 13:28. https://doi.org/10.1186/s13244-022-01170-2
5. Corrias G, Micheletti G, Barberini L et al (2022) Texture analysis imaging "what a clinical radiologist needs to know". Eur J Radiol 146:110055. https://doi.org/10.1016/j.ejrad.2021.110055
6. Yankeelov TE, Mankoff DA, Schwartz LH et al (2016) Quantitative imaging in cancer clinical trials. Clin Cancer Res 22:284–290. https://doi.org/10.1158/1078-0432.CCR-14-3336
7. Forghani R, Savadjiev P, Chatterjee A et al (2019) Radiomics and artificial intelligence for biomarker and prediction model development in oncology. Comput Struct Biotechnol J 17:995–1008. https://doi.org/10.1016/j.csbj.2019.07.001
8. Wu C, Lorenzo G, Hormuth DA 2nd et al (2022) Integrating mechanism-based modeling with biomedical imaging to build practical digital twins for clinical oncology. Biophys Rev 3:021304. https://doi.org/10.1063/5.0086789
9. Boehm KM, Khosravi P, Vanguri R et al (2022) Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer 22:114–126. https://doi.org/10.1038/s41568-021-00408-3
10. Bhinder B, Gilvary C, Madhukar NS, Elemento O (2021) Artificial intelligence in cancer research and precision medicine. Cancer Discov 11:900–915. https://doi.org/10.1158/2159-8290.CD-21-0090
11. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25:44–56. https://doi.org/10.1038/s41591-018-0300-7
12. Rundo L, Militello C, Vitabile S et al (2019) A survey on nature-inspired medical image analysis: a step further in biomedical data integration. Fundam Inform 171:345–365. https://doi.org/10.3233/FI-2020-1887
13. Nagendran M, Chen Y, Lovejoy CA et al (2020) Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 368:m689. https://doi.org/10.1136/bmj.m689
14. Rundo L, Pirrone R, Vitabile S et al (2020) Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. J Biomed Inf 108:103479. https://doi.org/10.1016/j.jbi.2020.103479
15. Makino T, Jastrzębski S, Oleszkiewicz W et al (2022) Differences between human and machine perception in medical diagnosis. Sci Rep 12:6877. https://doi.org/10.1038/s41598-022-10526-z
16. Shah P, Kendall F, Khozin S et al (2019) Artificial intelligence and machine learning in clinical development: a translational perspective. NPJ Digit Med 2:69. https://doi.org/10.1038/s41746-019-0148-3
17. Nanni L, Ghidoni S, Brahnam S (2017) Handcrafted vs. non-handcrafted features for computer vision classification. Pattern Recognit 71:158–172. https://doi.org/10.1016/j.patcog.2017.05.025
18. Zhang W, Guo Y, Jin Q (2023) Radiomics and its feature selection: a review. Symmetry 15:1834. https://doi.org/10.3390/sym15101834
19. Yamashita R, Nishio M, Do RKG, Togashi K (2018) Convolutional neural networks: an overview and application in radiology. Insights Imaging 9:611–629. https://doi.org/10.1007/s13244-018-0639-9
20. Hosny A, Aerts HJ, Mak RH (2019) Handcrafted versus deep learning radiomics for prediction of cancer therapy response. Lancet Digit Health 1:e106–e107. https://doi.org/10.1016/S2589-7500(19)30062-7
21. Guan H, Liu M (2022) Domain adaptation for medical image analysis: a survey. IEEE Trans Biomed Eng 69:1173–1185. https://doi.org/10.1109/TBME.2021.3117407
22. Shin H-C, Roth HR, Gao M et al (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35:1285–1298. https://doi.org/10.1109/TMI.2016.2528162
23. Tjoa E, Guan C (2021) A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 32:4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314
24. Castiglioni I, Rundo L, Codari M et al (2021) AI applications to medical images: from machine learning to deep learning. Phys Med 83:9–24. https://doi.org/10.1016/j.ejmp.2021.02.006
25. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: images are more than pictures, they are data. Radiology 278:563–577. https://doi.org/10.1148/radiol.2015151169
26. Kocak B, Baessler B, Bakas S et al (2023) Checklist for evaluation of radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. Insights Imaging 14:75. https://doi.org/10.1186/s13244-023-01415-8
27. van Timmeren JE, Cester D, Tanadini-Lang S et al (2020) Radiomics in medical imaging-"how-to" guide and critical reflection. Insights Imaging 11:91. https://doi.org/10.1186/s13244-020-00887-2
28. Aerts HJWL, Velazquez ER, Leijenaar RTH et al (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5:4006. https://doi.org/10.1038/ncomms5006
29. Zwanenburg A, Vallières M, Abdalah MA et al (2020) The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 295:328–338. https://doi.org/10.1148/radiol.2020191145
30. Santos DP, dos Santos DP, Dietzel M, Baessler B (2021) A decade of radiomics research: are images really data or just patterns in the noise? Eur Radiol 31:1–4. https://doi.org/10.1007/s00330-020-07108-w
31. Prinzi F, Militello C, Conti V, Vitabile S (2023) Impact of wavelet kernels on predictive capability of radiomic features: a case study on COVID-19 chest x-ray images. J Imaging 9:32. https://doi.org/10.3390/jimaging9020032
32. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J et al (2018) Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. Radiology 288:407–415. https://doi.org/10.1148/radiol.2018172361
33. Saha A, Yu X, Sahoo D, Mazurowski MA (2017) Effects of MRI scanner parameters on breast cancer radiomics. Expert Syst Appl 87:384–391. https://doi.org/10.1016/j.eswa.2017.06.029
34. Soleymani Y, Jahanshahi AR, Pourfarshid A, Khezerloo D (2022) Reproducibility assessment of radiomics features in various ultrasound scan settings and different scanner vendors. J Med Imaging Radiat Sci 53:664–671. https://doi.org/10.1016/j.jmir.2022.09.018
35. Scalco E, Belfatto A, Mastropietro A et al (2020) T2w-MRI signal normalization affects radiomics features reproducibility. Med Phys 47:1680–1691. https://doi.org/10.1002/mp.14038
36. Fornacon-Wood I, Mistry H, Ackermann CJ et al (2020) Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. Eur Radiol 30:6241–6250. https://doi.org/10.1007/s00330-020-06957-9
37. Whybra P, Zwanenburg A, Andrearczyk V et al (2024) The Image Biomarker Standardization Initiative: standardized convolutional filters for reproducible radiomics and enhanced clinical insights. Radiology 310:e231319. https://doi.org/10.1148/radiol.231319
38. Cattell R, Chen S, Huang C (2019) Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. Vis Comput Ind Biomed Art 2:19. https://doi.org/10.1186/s42492-019-0025-6
39. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86:420–428. https://doi.org/10.1037/0033-2909.86.2.420
40. Přibil J, Přibilová A, Frollo I (2019) Analysis of the influence of different settings of scan sequence parameters on vibration and noise generated in the open-air MRI scanning area. Sensors (Basel) 19:4198. https://doi.org/10.3390/s19194198
41. Gravina M, Marrone S, Docimo L et al (2022) Leveraging CycleGAN in lung CT sinogram-free kernel conversion. In: Proceedings of the 21st International Conference on Image Analysis and Processing—ICIAP 2022. Springer, Heidelberg, pp 100–110. https://doi.org/10.1007/978-3-031-06427-2_9

42. Yang S, Kim EY, Ye JC (2021) Continuous conversion of CT kernel using switchable CycleGAN with AdaIN. IEEE Trans Med Imaging 40:3015–3029. https://doi.org/10.1109/TMI.2021.3077615
43. Lei Y, Harms J, Wang T et al (2019) MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. Med Phys 46:3565–3581. https://doi.org/10.1002/mp.13617
44. Svoboda D, Burgos N, Wolterink JM, Zhao C (2021) Simulation and synthesis in medical imaging. In: Proceedings of the 6th International Workshop, SASHIMI 2021, held in conjunction with MICCAI 2021, Strasbourg, 27 September 2021. Springer. https://doi.org/10.1007/978-3-030-87592-3
45. Ozbey M, Dalmaz O, Dar SUH et al (2023) Unsupervised medical image translation with adversarial diffusion models. IEEE Trans Med Imaging 42:3524–3539. https://doi.org/10.1109/TMI.2023.3290149
46. Kazerouni A, Aghdam EK, Heidari M et al (2023) Diffusion models in medical imaging: a comprehensive survey. Med Image Anal 88:102846. https://doi.org/10.1016/j.media.2023.102846
47. Mahon RN, Ghita M, Hugo GD, Weiss E (2020) ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. Phys Med Biol 65:015010. https://doi.org/10.1088/1361-6560/ab6177
48. Horng H, Singh A, Yousefi B et al (2022) Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects. Sci Rep 12:4493. https://doi.org/10.1038/s41598-022-08412-9
49. Ardakani AA, Nathalie JB, Ciaccio EJ, Rajendra Acharya U (2022) Interpretation of radiomics features—a pictorial review. Comput Methods Prog Biomed 215:106609. https://doi.org/10.1016/j.cmpb.2021.106609
50. Tomaszewski MR, Gillies RJ (2021) The biological meaning of radiomic features. Radiology 299:E256. https://doi.org/10.1148/radiol.2021202553
51. Wei P (2021) Radiomics, deep learning and early diagnosis in oncology. Emerg Top Life Sci 5:829–835. https://doi.org/10.1042/ETLS20210218
52. Pesapane F, Codari M, Sardanelli F (2018) Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2:35. https://doi.org/10.1186/s41747-018-0061-6
53. Ferreira MD, Corrêa DC, Nonato LG, de Mello RF (2018) Designing architectures of convolutional neural networks to solve practical problems. Expert Syst Appl 94:205–217. https://doi.org/10.1016/j.eswa.2017.10.052
54. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
55. Khan S, Rahmani H, Shah SAA, Bennamoun M (2022) Features and classifiers. In: A guide to convolutional neural networks for computer vision. Synthesis Lectures on Computer Vision. Springer, Cham. https://doi.org/10.1007/978-3-031-01821-3_2
56. Cui S, Tseng H-H, Pakela J et al (2020) Introduction to machine and deep learning for medical physicists. Med Phys 47:e127–e147. https://doi.org/10.1002/mp.14140
57. Litjens G, Kooi T, Bejnordi BE et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005
58. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35:1798–1828. https://doi.org/10.1109/TPAMI.2013.50
59. Anuradha T, Tigadi A, Ravikumar M et al (2022) Feature extraction and representation learning via deep neural network. In: Computer networks, Big Data and IoT. Springer Nature, Singapore, pp 551–564. https://doi.org/10.1007/978-981-19-0898-9_44
60. Venkataramani R, Ravishankar H, Anamandra S (2019) Towards continuous domain adaptation for medical imaging. In: Proceedings of 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice. https://doi.org/10.1109/ISBI.2019.8759268
61. Tajbakhsh N, Shin JY, Gurudu SR et al (2016) Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging 35:1299–1312. https://doi.org/10.1109/TMI.2016.2535302
62. Parmar C, Grossmann P, Bussink J et al (2015) Machine learning methods for quantitative radiomic biomarkers. Sci Rep 5:1–11. https://doi.org/10.1038/srep13087
63. Sun P, Wang D, Mok VC, Shi L (2019) Comparison of feature selection methods and machine learning classifiers for radiomics analysis in glioma grading. IEEE Access 7:102010–102020. https://doi.org/10.1109/ACCESS.2019.2928975
64. Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, New York, NY. https://doi.org/10.1007/978-0-387-21606-5
65. Demircioğlu A (2022) Benchmarking feature selection methods in radiomics. Invest Radiol 57:433–443. https://doi.org/10.1097/RLI.0000000000000855
66. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40:16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024
67. Dasarathy BV, Holder EB (1991) Image characterizations based on joint gray level—run length distributions. Pattern Recognit Lett 12:497–502. https://doi.org/10.1016/0167-8655(91)80014-2
68. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97:273–324. https://doi.org/10.1016/S0004-3702(97)00043-X
69. Varoquaux G, Cheplygina V (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med 5:48. https://doi.org/10.1038/s41746-022-00592-y
70. Khushi M, Shaukat K, Alam TM et al (2021) A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access 9:109960–109975. https://doi.org/10.1109/ACCESS.2021.3102399
71. Avanzo M, Wei L, Stancanello J et al (2020) Machine and deep learning methods for radiomics. Med Phys 47:e185–e202. https://doi.org/10.1002/mp.13678
72. Roffo G, Melzi S, Castellani U, Vinciarelli A (2017) Infinite latent feature selection: a probabilistic latent graph-based ranking approach. In: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Venice, Italy. pp 1398–1406. https://openaccess.thecvf.com/content_ICCV_2017/papers/Roffo_Infinite_Latent_Feature_ICCV_2017_paper.pdf
73. Militello C, Rundo L, Dimarco M et al (2022) 3D DCE-MRI radiomic analysis for malignant lesion prediction in breast cancer patients. Acad Radiol 29:830–840. https://doi.org/10.1016/j.acra.2021.08.024
74. Nahm FS (2022) Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol 75:25–36. https://doi.org/10.4097/kja.21209
75. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Lond Edinb Dubl Philos Mag J Sci 2:559–572. https://doi.org/10.1080/14786440109462720
76. Hinton GE, Roweis S (2002) Stochastic neighbor embedding. Adv Neural Inf Process Syst 15:857–864. https://cs.nyu.edu/~roweis/papers/sne_final.pdf
77. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf
78. McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426
79. Bloom V, Argyriou V, Makris D (2017) Linear latent low dimensional space for online early action recognition and prediction. Pattern Recognit 72:532–547. https://doi.org/10.1016/j.patcog.2017.07.003
80. Recanatesi S, Farrell M, Lajoie G et al (2021) Predictive learning as a network mechanism for extracting low-dimensional latent space representations. Nat Commun 12:1417. https://doi.org/10.1038/s41467-021-21696-1
81. Kaissis GA, Makowski MR, Rückert D, Braren RF (2020) Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2:305–311. https://doi.org/10.1038/s42256-020-0186-1
82. Rieke N, Hancox J, Li W et al (2020) The future of digital health with federated learning. NPJ Digit Med 3:119. https://doi.org/10.1038/s41746-020-00323-1
83. Nan Y, Ser JD, Walsh S et al (2022) Data harmonisation for information fusion in digital healthcare: a state-of-the-art systematic review, meta-analysis and future research directions. Inf Fusion 82:99–122. https://doi.org/10.1016/j.inffus.2022.01.001
84. Dhade P, Shirke P (2024) Federated learning for healthcare: a comprehensive review. Eng Proc 59:230. https://doi.org/10.3390/engproc2023059230
85. Clusmann J, Kolbinger FR, Muti HS et al (2023) The future landscape of large language models in medicine. Commun Med 3:1–8. https://doi.org/10.1038/s43856-023-00370-1

86. Bajaj S, Gandhi D, Nayar D (2023) Potential applications and impact of ChatGPT in radiology. Acad Radiol 31:1256–1261. https://doi.org/10.1016/j.acra.2023.08.039
87. Li D, Gupta K, Chong J (2023) Evaluating diagnostic performance of ChatGPT in radiology: delving into methods. Radiology 308:e232082. https://doi.org/10.1148/radiol.232082
88. Akinci D'Antonoli T, Stanzione A, Bluethgen C et al (2024) Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol 30:80–90. https://doi.org/10.5167/uzh-239764
89. Hager P, Jungmann F, Holland R et al (2024) Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. https://doi.org/10.1038/s41591-024-03097-1
90. Harang R, Sanders H (2023) Catastrophic forgetting in the context of model updates. Preprint at https://doi.org/10.48550/arXiv.2306.10181
91. Kirkpatrick J, Pascanu R, Rabinowitz N et al (2017) Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci U S A 114:3521–3526. https://doi.org/10.1073/pnas.1611835114
92. Ren W, Li X, Wang L et al (2024) Analyzing and reducing catastrophic forgetting in parameter efficient tuning. Preprint at https://doi.org/10.48550/arXiv.2402.18865
93. Castellano G, Bonilha L, Li LM, Cendes F (2004) Texture analysis of medical images. Clin Radiol 59:1061–1069. https://doi.org/10.1016/j.crad.2004.07.008
94. Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Trans Syst Man Cybern 3:610–621. https://doi.org/10.1109/TSMC.1973.4309314
95. Haralick RM (1979) Statistical and structural approaches to texture. Proc IEEE 67:786–804. https://doi.org/10.1109/PROC.1979.11328
96. Galloway MM (1975) Texture analysis using gray level run lengths. Comput Graph Image Process 4:172–179. https://ui.adsabs.harvard.edu/abs/1974STIN…7518555G
97. Chu A, Sehgal CM, Greenleaf JF (1990) Use of gray value distribution of run lengths for texture analysis. Pattern Recognit Lett 11:415–419. https://doi.org/10.1016/0167-8655(90)90112-F
98. Sun C, Wee WG (1982) Neighboring gray level dependence matrix for texture classification. Comput Graph Image Process 20:297. https://doi.org/10.1016/0734-189X(83)90032-4
99. Amadasun M, King R (1989) Textural features corresponding to textural properties. IEEE Trans Syst Man Cybern 19:1264–1274. https://doi.org/10.1109/21.44046
100. Thibault G, Angulo J, Meyer F (2014) Advanced statistical matrices for texture characterization: application to cell classification. IEEE Trans Biomed Eng 61:630–637. https://doi.org/10.1109/TBME.2013.2284600
101. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego. https://doi.org/10.1109/CVPR.2005.177
102. Ojala T, Pietikäinen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distributions. Pattern Recognit 29:51–59. https://doi.org/10.1016/0031-3203(95)00067-4
103. Gabor D (1946) Theory of communication. Part 1: The analysis of information. J Inst Electr Eng Radiol Commun Eng 93:429–441. https://doi.org/10.1049/ji-3-2.1946.0074
104. Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 11:674–693. https://doi.org/10.1109/34.192463
105. Soda P, D'Amico NC, Tessadori J et al (2021) AIforCOVID: predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study. Med Image Anal 74:102216. https://doi.org/10.1016/j.media.2021.102216

Publisher's Note