Diagn Interv Radiol 2024; DOI: 10.4274/dir.2024.242719

ARTIFICIAL INTELLIGENCE AND INFORMATICS

INVITED REVIEW

Reproducibility and interpretability in radiomics: a critical assessment

Aydın Demircioğlu

University Hospital Essen, Institute of Diagnostic and Interventional Radiology and Neuroradiology, Essen, Germany

ABSTRACT

Radiomics aims to improve clinical decision making through the use of radiological imaging. However, the field is challenged by reproducibility issues due to variability in imaging and subsequent statistical analysis, which particularly affects the interpretability of the model. In fact, radiomics extracts many highly correlated features that, combined with the small sample sizes often found in radiomics studies, result in high-dimensional datasets. These datasets, which are characterized by containing more features than samples, have different statistical properties than other datasets, thereby complicating their training by machine learning and deep learning methods. This review critically examines the challenges of both reproducibility issues and interpretability, beginning with an overview of the radiomics pipeline, followed by a discussion of the imaging and statistical reproducibility issues. It further highlights how limited model interpretability hinders clinical translation. The discussion concludes that these challenges could be mitigated by following best practices and by creating large, representative, and publicly available datasets.

KEYWORDS

Computer-aided detection, deep learning, interpretability, radiomics, reproducibility, statistics

Radiomics is a recent field that uses "an automated high-throughput extraction of large amounts of quantitative features of medical images."[1-3] The method "converts imaging data into a high dimensional mineable feature space using a large number of automatically extracted data-characterization algorithms."[4]

The above definition may seem complex, but it can be succinctly summarized. Similar to how clinical routine involves characterizing a patient using parameters such as age, weight, and hemoglobin levels, radiological images can be analyzed to extract analogous parameters (also called features) that ideally describe the pathology of interest. For example, in the case of a tumor lesion, features such as its volume and diameter can be measured. A critical aspect of radiomics is the extraction of not only morphological features but also the distribution of intensity and texture. This includes, for instance, assessing whether the lesion has high brightness and a homogeneous or coarse texture, and identifying the presence of bright spots. Radiomics involves the extraction of hundreds to thousands of such features to accurately represent the lesion. These features are subsequently used to train a classifier, that, based on the characteristics of a new lesion, can determine, for example, whether the lesion is benign.

The main expectation of radiomics is that these features can serve as surrogates for biomarkers, and thus aid clinical decision making. Radiological imaging could reflect the underlying biological processes, allowing for indirect conclusions. For example, while necrotic cells are not directly observable in computed tomography (CT) scans, their presence may result in the appearance of a hypodense lesion (Figure 1). Thus, measuring the overall intensity of a lesion could be used as an indicator of cell necrosis.

Although radiomics as a field only emerged in the 2010s,[1,5] the idea can be traced back much further. In a seminal paper published in 1978, Harlow et al.[6] introduced concepts that are strikingly similar. Later, specifically in the 1990s, similar techniques were introduced as texture analysis.[7] This is no coincidence, since the underlying idea of applying machine learn-

Corresponding author: Aydın Demircioğlu

E-mail: aydin.demircioglu@uk-essen.de

Received 02 July 2024; revision requested 13 August 2024; accepted 12 September 2024.

Epub: 21.10.2024

Publication date: 08.07.2025

DOI: 10.4274/dir.2024.242719

You may cite this article as: Demircioğlu A. Reproducibility and interpretability in radiomics: a critical assessment. *Diagn Interv Radiol.* 2025;31(4):321-328.

ing to imaging is the same and dates back to the 1960s.

The primary purpose of radiomics is to support clinical decisions. Ideally, the extracted features provide insights that humans cannot see or systematically process, allowing clinicians to answer questions using this hidden information. Radiomics has also been used to non-invasively identify genetic alterations or gene expression patterns that can be used to predict the outcome or survival risk of patients with cancer.[8-10]

In this review, the basic concepts of radiomics are first introduced, followed by a detailed discussion of the two major reproducibility issues that persist in the current field. Subsequently, radiomics based on deep neural networks is briefly outlined and the issues involved in their application examined. Finally, strategies for avoiding these issues are discussed.

The radiomics pipeline

As with any study, the first step in a radiomics model is to define patient cohorts, applying reasonable inclusion and exclusion criteria that reflect the target population, and defining an outcome of clinical interest.

The application of radiomics to data is technical but relatively straightforward (Figure 2).[11] Images are first acquired and the region of interest (ROI) is segmented. This can be a tumor lesion or an entire organ, such as the whole prostate. The ROI plays a critical role in directing the analysis to relevant areas, thereby preventing other unrelated regions from potentially confounding the analysis.

The images are then pre-processed depending on the use case. For example, magnetic resonance imaging may require a normalization step, and CT may be thresholded to a Hounsfield units range of interest. In addition, preprocessing filters are applied. For instance, smoothing filters can reduce noise that may adversely affect features, whereas wavelet filters can decompose the image into high-frequency and low-frequency components that may carry different information, aiding subsequent analysis.

Next, features are extracted from the ROI. This is a central step, and there are three main types of generic features that are extracted: morphological features, such as volume or sphericity; intensity features, which measure the distribution of values, such as mean brightness; and texture features that reflect the co-occurrence of intensity values.

However, feature extraction will often generate large numbers of features, and many of them will be irrelevant (i.e., they will not help to solve the problem). Many will also be redundant, that is, their information is already present in other features. Therefore, a feature selection step is applied that retains only the relevant features; for example, a t-test can be used to filter out those that are not significant.

These features are then fed into a classifier, which functions in terms of making a prediction after receiving a set of features. This classifier is trained on the data using machine learning techniques. In other words, following the input of data, the algorithm identifies relevant patterns to make accurate predictions on new data. This model can then be tested and applied to new data, such as routine clinical data.

Main points

- Radiomics is impeded by imaging and statistical reproducibility issues.
- Machine and deep learning modeling are complicated and require extensive validation.
- Radiomic features found to be predictive in modeling often do not correspond to biomarkers due to high correlation, limiting their interpretability.
- Standardization practices and larger, more diverse datasets are important to improve reproducibility.

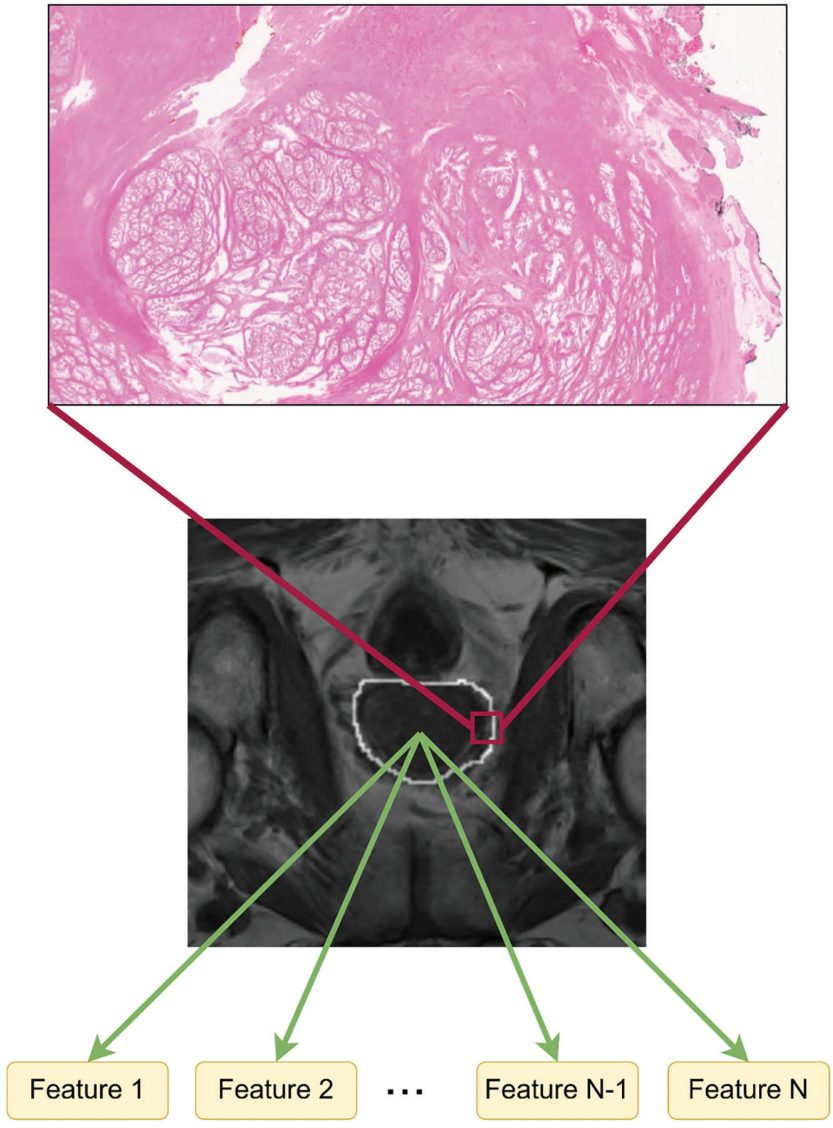

Figure 1. Radiomics aims to identify biomarkers by measuring them indirectly through radiological imaging. Much of the information in the pathological scan (top) is lost in the radiological image (bottom). Features are extracted from the segmented region-of-interest to recover the information of interest (the pathology image is part of the PROSTATE-MRI dataset).[75] MRI, magnetic resonance imaging.

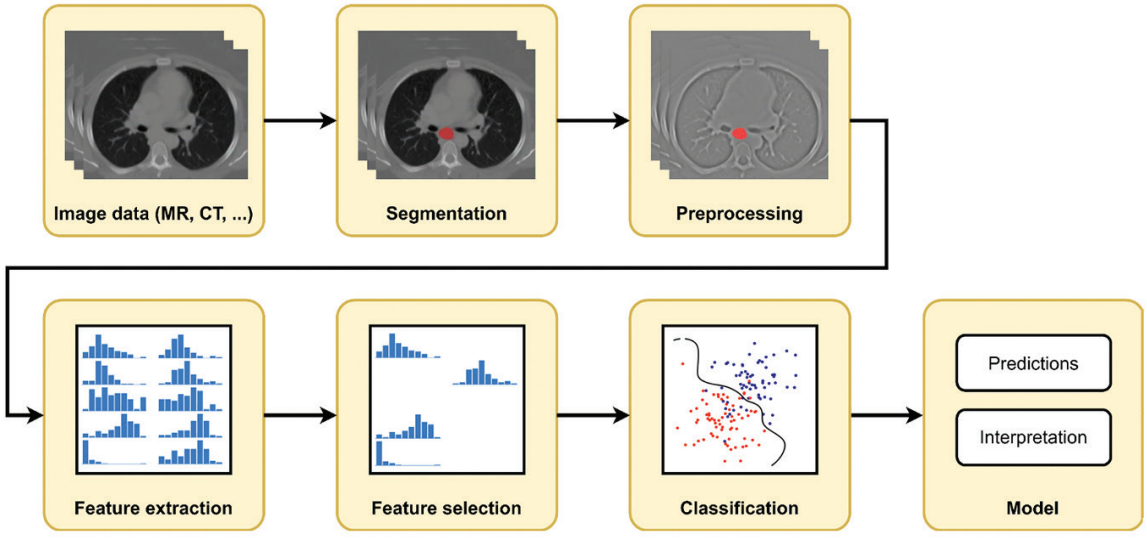

Figure 2. Brief overview of the radiomics pipeline. MR, magnetic resonance; CT, computed tomography.

The radiomics pipeline appears pretty straightforward, but in each step, good practices must be maintained to avoid biased or false-positive results.[12]

Reproducibility issues

Although the pipeline may seem fairly rigid, the key issue is reproducibility. This term describes the requirement that similar findings should be observed when conditions do not change significantly. For example, scanning the same patient twice within a very short time frame should yield similar radiomic features and lead to similar predictions. Non-reproducible studies are essentially random and erratic and cannot be trusted. They may also lead to false positives, which would prevent clinical use.

Reproducibility in radiomics can be divided into two areas: imaging reproducibility and statistical reproducibility. The term "imaging reproducibility" refers to the acquisition of scans and the extraction of features, whereas "statistical reproducibility" refers to modeling using machine learning. Of course, if the imaging is not reproducible, no modeling can correct it (following the well-known "garbage in, garbage out" rule).[13,14] Nonetheless, the focus will be mainly on statistical reproducibility.

Imaging reproducibility

Imaging reproducibility refers to issues in the acquisition process resulting from variations in imaging parameters and techniques, vendor differences, and similar factors.[15] Since radiomic features are extracted from the acquired images, parameters such as voxel size and reconstruction techniques can have a significant impact on these features.[16,17] The effect is also non-linear, which can render images highly non-reproducible.[18] Post-hoc harmonization can mitigate the problem, but only to a limited extent.[19,20]

Even if the imaging were reproducible, the segmentations are usually sensitive to intra- and inter-rater variability, and these differences can also have a large impact on the extracted features,[21] making them partially non-reproducible. The same is true for the definition of the features themselves. Even simple features, such as sphericity, can show variations depending on the formulas used to calculate them. Accordingly, the Image Biomarker Standardisation Initiative (IBSI) was launched to standardize these features and assess their reproducibility.[22] However, not all software programs are IBSI-compliant, and even the standardized features may still exhibit some differences.[23]

Another source of variability is the use of preprocessing filters. Although standardization has recently been considered by the IBSI,[24] it is unknown whether preprocessing helps at all, and if so, which filters should be applied. Therefore, these preprocessing filters are applied in parallel to increase the predictive power of the resulting features.[25] However, this leads to statistical problems.

Statistical reproducibility

The data generated will often have two characteristics that distinguish it from many other datasets: it will be high-dimensional, meaning that there are more features than samples, and it will be highly correlated. In radiomics, there are two main reasons for this. First, the total sample size is often limited due to the time and resources required for annotation, the rarity of the disease in question, or privacy concerns. Second, the numerous preprocessing filters extract information that is highly similar. For example, two levels of smoothing will produce features that are very alike. This results in the generation of highly correlated features.

The presence of such data presents significant challenges, as the search for predictive features and patterns becomes exponentially more difficult and resembles "finding a needle in a haystack."[26] Therefore, the risk of identifying spurious patterns and producing false-positive results is significantly increased in such data. While methods such as regularization can help overcome this problem, the issue remains unresolved.

Therefore, radiomics often employs a feature selection step, where the goal is to retain only the relevant features and remove all others, thereby reducing the dimensionality of the data. However, several methods of varying complexity are currently in use.[11,27,28] Simpler methods, such as Spearman correlation or t-tests, typically operate by considering each variable on its own. These methods are computationally efficient but may overlook dependencies between variables, potentially leading to suboptimal feature selection. More complex methods, such as the least absolute shrinkage and selection operator method,[29] the minimum redundance maximum relevancy method,[30] or the Boruta method,[31] are able to account for such dependencies but are more computationally demanding. While it may be intuitive to assume that more complex methods perform better, it has been shown that for many data-

sets, the differences may not be significant. However, simpler methods tend to be more robust, and therefore more reproducible.[27] In addition, many of the feature selection methods do not select relevant features but merely score them, leaving open the decision regarding how many of the highest-scoring features to retain, which reduces their reproducibility.

Accordingly, feature selection is not a complete solution to the problem since the task of dealing with the high-dimensional space is merely transferred from the classifier. Feature selection is also subject to failure and may even underperform, especially given the inherent instability of selection methods and their dependence on the specific data sample.[27] For example, the removal of a few samples can have a significant impact on the set of features considered relevant.

Subsequent classifiers are also affected by high dimensionality, either directly or indirectly, if irrelevant features have been selected. Furthermore, many classifiers make assumptions about the data that may not be true, regardless of whether feature selection has been applied. These assumptions are often controlled by hyperparameters; for example, a regularization variable may reflect the amount of noise present in the data. Therefore, the only option is to test many different parameters, which is extremely expensive in terms of computational resources. As a result, studies only test a limited number of parameters, and it remains unclear whether a significantly more effective model could have been obtained by optimizing the hyperparameters.

Validation issues

Any model requires extensive testing, the main reason for this being that models could either memorize the data or find spurious instead of predictive patterns. Such a model would perform well during training, but worse on test data and would not generalize. This problem is called overfitting, and the risk is higher for high-dimensional data, where more patterns can fit the given data.

To avoid this problem, validation is performed first. Unlike testing, validation is mainly used for model selection, specifically to determine good values for the hyperparameters, or to identify which feature selection or classifier method performs better on the given data. Ideally, validation should be performed on a second independent dataset, but alternatively, a portion of the data can be set aside. Certain common schemes are often employed in radiomics, including simple splitting, cross-validation, and bootstrapping. In simple splitting, a portion of the data (e.g., 70%) is used for training, whereas the remainder is used exclusively for validation. While this method is conceptually simple and computationally fast, it does not utilize all available data for training. Additionally, the results can be highly dependent on the specific split, leading to potential variability; that is, there is a risk that results may be good, or bad, by chance. To mitigate this, the method can be repeated several times and the results averaged. Cross-validation provides a more systematic approach by splitting the data into *k* subsets and iteratively training on *k*-1 subsets while using the remaining subset for validation. Although computationally more expensive, this method ensures that all data is used for both training and validation, providing a more reliable estimate of the performance. Nested cross-validation further refines this by applying cross-validation twice: once to the entire data for performance estimation and once on the training data for hyperparameter tuning. This scheme provides an unbiased evaluation and is considered a gold standard. Bootstrapping, on the other hand, uses resampling with replacement to create training and validation sets. Since samples can occur multiple times in the training set, this approach simulates different weights for each sample and can thus lead to better estimates. However, to obtain these estimates, a large number of repetitions (e.g., 1,000) is generally required, making it computationally highly expensive.

However, in all cases, the golden rule of machine learning must be followed: training and test sets must be kept strictly separate. Failure to follow this rule will lead to data leakage, meaning that the classifier has already seen some aspects of the test data and could adapt to it, leading to false positives.[32,33]

Another issue is the variability of the data. Choosing a homogeneous cohort (e.g., from a single scanner) increases the likelihood of obtaining a working model since the predictive patterns seen during training are likely to be present in the test data. At the same time, however, the model will be highly specific and may not generalize beyond the collected data. The opposite, collecting heterogeneous data, is also critical, because the classifier may not be able to identify any predictive patterns at all, especially with small sample sizes, and there will be no relevant model to test. However, if such a model is successful, its clinical applicability will be much higher, which is the ultimate goal.[34]

Deep radiomics

Deep learning has recently shown great success in other fields,[35] and it is natural to apply deep learning to radiomics. Deep learning is based on artificial neural networks, which, in a simplistic way, try to mimic the human brain, and date back to the early days of machine learning in the 1950s. Conceptually, in the simplest case, a network consists of multiple layers, each of which can be understood as a feature generation step. Layer by layer, the input is transformed into the desired output, and the training data is used to determine the parameters of the layers (Figure 3).

Applying deep learning to radiomics, which is termed deep radiomics, can, in contrast to the generic radiomics discussed above, mitigate two major drawbacks. First, it can potentially reduce the need for segmentation because the network can, at least potentially, determine the ROI itself. Equally important, the network can extract optimal features that are specific to the problem at hand. It can also consider more global features of the data, whereas most generic features are based on local textures. Both can lead to models that perform much better than generic models. While deep learning has only recently gained importance, neural networks have been applied to radiological data since the 1990s.[36,37]

Issues with deep radiomics

Deep radiomics does not magically bypass the reproducibility problems. For example, changes in acquisition parameters have been shown to have a strong effect on predictive performance, thus affecting generalizability.[38] Much is unknown about the stability of deep radiomics models, such as whether a different training sample will yield different features, or whether features from different networks are highly correlated. Robustness to image noise and slightly different segmentations has also not been systematically investigated, which is complicated by the fact that many different architectures exist.

Sample size is an even bigger issue in deep radiomics. Learning directly from data usually requires many more samples to be successful.[39] As a result, deep radiomics is currently not as successful as it could be.

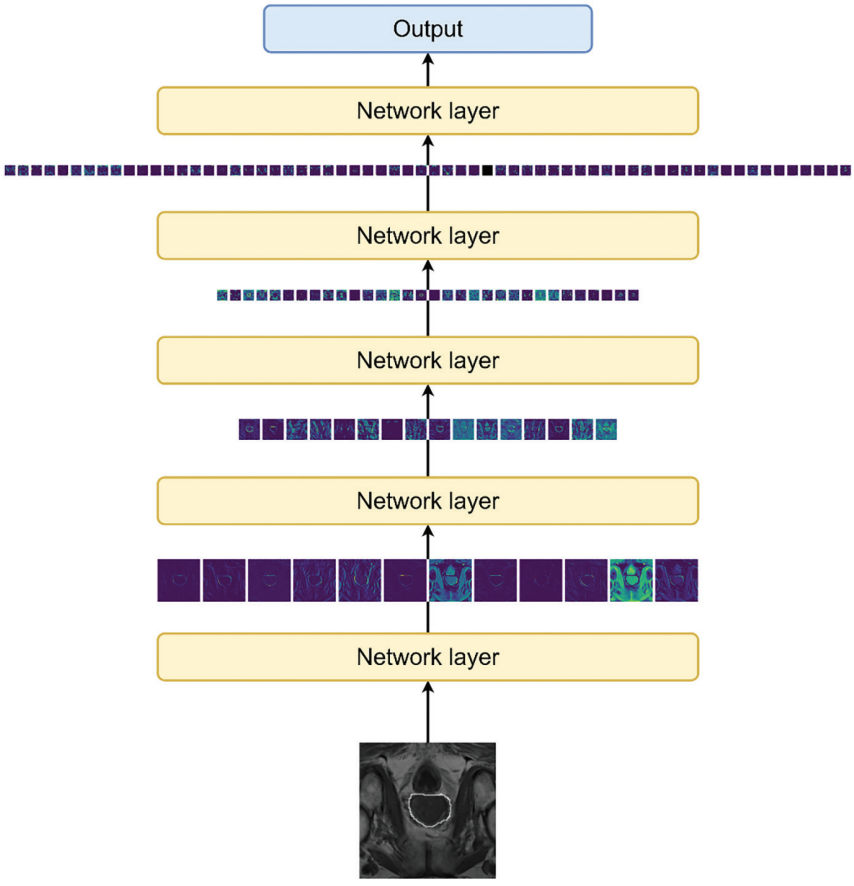

Figure 3. In simple terms, the network can be thought of as a set of layers that transform an input image into a set of output images. Each layer of the network has many parameters that are optimized using the training data. Networks usually do not use segmentation, but can be modified to use it. The network can be used as a feature generator by extracting features from the output of an appropriate layer. For example, in the figure, each of the 64 small images output by the second-to-last layer at the top could be averaged, resulting in 64 numerical features for the given input.

Consequently, several mitigation strategies have been developed.[40,41] However, they all have their own drawbacks. For example, studies often resort to using image slices for training, which not only increases the sample size but also allows for the use of smaller networks.[42,43] Nonetheless, this approach partially loses the spatial information, which reduces the potential benefit.

A more common strategy is transfer learning. Here, the network is first trained on a dataset from another domain, most commonly ImageNet, a collection of photographs.[44] This pre-trained network is then fine-tuned (i.e., it is trained on the radiomic data, often at lower learning rates) to slightly adjust the network. This approach can work because there is a remarkable similarity between the low-level features of the human eye and the network; at lower levels, both appear to operate with filters comparable to Gabor filters.[39] Thus, fine-tuning can focus on training the higher layers and performing better with fewer samples. However, the use of non-medical data for pre-training is again suboptimal, and larger medical data corpora have been introduced only recently, although the extent to which these can help in radiomics remains unclear, as they are usually far smaller than ImageNet.[45]

Since training a deep network involves many hyperparameters (e.g., learning rate, learning schedule, choice of loss function) and can be relatively complicated, another alternative is to bypass any training and instead use only pre-trained networks as feature extractors (Figure 3),[46] which allows more versatile classifiers, such as boosting, to perform better, especially with smaller sample sizes.[47] However, since no training is performed in this approach, the disadvantage is again that the features may be less optimal, although fusing them with generic radiomics can still prove helpful.[48,49]

Finally, the hope that deep radiomics can dispense with segmentation may be in vain due to the small sample size. In addition, without a proper validation method, deep radiomics is also prone to bias due to over-engineering. In fact, a recent review found no clear advantage of deep radiomics.[50]

Interpretability issues

A key point in radiomics is to identify features that can potentially serve as biomarkers, just as the volume of a lesion indicates its malignancy. However, radiomics attempts to establish such a correspondence "in reverse," using the coarser and noisier radiological images, where much information is already lost during acquisition. Radiomics seeks to capture the underlying information by making multiple measurements (in the form of different features). These are often correlated, as they can be understood as noisy and incomplete versions of the inaccessible information. There is no guarantee that the information can be recovered from the extracted features, nor that the observed predictivity of a feature actually corresponds to a biomarker.

Given a set of features, radiomics can only identify those that are statistically associated with the outcome. Such an association is not causal and could only be the basis of a subsequent statistically sound test. This problem is exacerbated by the high-dimensionality of the data, where the intuition from the low-dimensional setting that features have a clear meaning and their importance can be easily measured fails.[51] In fact, the very concept of distance becomes somewhat incomprehensible in higher dimensions, often termed the curse of dimensions, and is demonstrated by the fact that in higher dimensions, most of the volume of a unit sphere is near its surface.[52]

In fact, the use of feature importance as a surrogate has been shown to be questionable because essentially every step in the radiomics pipeline affects the importance of features in the resulting model. Even seemingly unimportant preprocessing steps, such as the choice of discretization method[23] and data normalization, which is performed to obtain the data on a uniform scale, can strongly influence the set of features and thus the interpretability.[53] This influence is more evident in the feature selection step, where different methods will emphasize different aspects and thus gain different importance.[27] Not only does the subsequent classifier affect interpretation but the selection of the final model can also have a great impact, as often several models will perform very

closely but will select different sets of features as important.[51,54] In a systematic review, Tohidinezhad et al.[55] identified 23 models that predict the effect of radiation on brain health. None of these models used exactly the same features, and the models differed widely in the factors that were significantly associated with outcome.

Moreover, even if such an identification were possible, most radiomic features are not interpretable by themselves. For example, it is unclear what semantic meaning a feature such as wavelet-LHL_glrlm_GrayLevelNonUniformityNormalized carries, and how to see the difference from a highly correlated feature that is slightly less predictive. It is unlikely that a radiologist would be able to relate the measured values of such a feature to the scan. Feature maps may be helpful for visualization,[56] but they are currently only a tool and cannot be used to base an interpretation on. In addition, radiomic models are rarely based on a single feature, and a meaningful interpretation of a model using multiple features is barely possible. Paradoxically, radiomics was invented precisely because humans cannot describe textural patterns well.

The potential for highly correlated features to cause interpretation problems is illustrated by a recent study by Welch et al.,[57] who reexamined the model that Aerts et al.[4] used in their seminal work on patients with non-small cell lung cancer. The authors showed that volume alone is as predictive as the radiomic model, and moreover, that three of the four texture features found by Aerts et al.[4] are highly correlated with volume.

Recently, post-hoc interpretations, such as Explainable AI (XAI) methods, have been applied.[58] However, these are also problematic. Since there are several different XAI methods, it is likely that the resulting meanings will also differ.[59] Alternatively, explainable classifiers could be used, which generally involves a trade-off between the complexity (and thus interpretability) of the classifier and its predictive performance.[60] However, even if these methods are successful, they only address the classifier and do not mitigate the problems in the overall pipeline.

The situation is similar for deep radiomics. While the pipeline itself is less complex, training is more difficult, and there are many more choices regarding the architecture. It is highly likely that different choices will lead to vastly different features. In addition, the deep features do not have a mathematical formula, making any direct interpretation difficult. To remedy this situation, Cho et al.[61] correlated deep features with radiomic features. However, since radiomic features are not fully interpretable by themselves, this approach is limited in scope.

Discussion

Currently, radiomics suffers from both imaging and statistical reproducibility issues, both of which affect the interpretability and applicability of the models. This affects the entire radiomics pipeline, and even feature normalization can lead to reproducibility issues.

Neither of these problems can be easily avoided. Image reproducibility could possibly be mitigated by strict standardization of imaging protocols, but this is all but impossible to implement in practice across multiple centers. Statistical reproducibility is also not easily mitigated. Methodological differences aside, different research groups will often reach different conclusions given the same data.[62] Although such studies have not been conducted in radiomics, the impact is expected to be even greater, as there is generally less code and data sharing in the health domain.[63]

One major problem is small sample sizes. Radiomics studies need to include larger and more diverse datasets to have a chance of success. This is illustrated by current models that use deep learning to diagnose chest X-rays, or mammograms that have been shown to perform especially well.[64,65] These models are often trained on datasets that reach tens of thousands of scans. However, they are not radiomic in the sense that they do not require segmentations. The abundance of data makes segmentations unnecessary, as the network can identify the relevant regions on its own. Although it is virtually impossible to obtain such large sample sizes for rare cancers, more data would potentially reduce the dimensionality of the data and thus increase reproducibility. Nonetheless, radiomics seems to have made no progress since the seminal work of Harlow et al.[6] in 1976, where sample sizes of around 300 are reported. Small sample sizes are generally unable to reflect heterogeneity. This is even true for within-patient heterogeneity. For example, suppose two features are measured in a single patient at two time points, as in a test-retest scenario, and their sum is predictive. Then, the two features may vary greatly between the two time points such that neither is reproducible; but provided their sum remains the same, this would not pose any problem for their predictive value. However, if the model was not trained on such data, it would not find that pattern and would fail on new data. Nevertheless, large sample sizes are useless if the images do not carry the necessary information and such predictive patterns do not exist. Hence more data is not always helpful.

Non-reproducible studies may also result from a failure to follow best practices, which can be ensured by adhering to proper guidelines.[66,67] For example, the study must be described in full detail in a manner that enables replication by others. Code should always be shared, and data should be shared if possible. Best practices encompass every step of the study; for example, it must be ensured that the data selection is appropriate and unbiased relative to the study's objective.[12,68] The outcome should also be compared with current standards where applicable, for example, if a clinical scoring system is in current use (e.g., the Prostate Imaging Reporting and Data System), the radiomics model should be compared against it.[69] Statistical tests (e.g., permutation tests) can be used to ensure that the resulting model is different from a random guess, which is crucial when sample sizes are small. While statistical significance should be computed, the clinical significance should also be considered to evaluate the impact of the model. Furthermore, the overall study design must be methodologically sound to avoid reporting false-positive results. In addition, reporting must be clear and complete to ensure reproducibility.[70]

In a seminal paper, Ioannidis argued that around 60% of all medical studies contain false-positive results.[71] Studies with such obvious false positives should therefore be retracted, but this almost never happens in radiomics. On the contrary, such studies are frequently cited.[72] In addition, methodologically correct studies will fare relatively worse and may appear as "negative" studies that may not be considered for publication.[73] To mitigate this, a far more rigorous review process with mandatory code or data sharing would be required, as it could help in identifying potentially biased results before their publication. Currently, such studies are often only identified following publication, making it difficult to address the issue. Ensuring that publications rigorously follow reporting guidelines could be another way to reduce the problem.[66,67,70]

It is easy to overlook the fact that image processing has gone through a similar evo-

lution in the past. The field started with the manual extraction of many features (which is the origin of the texture features used today), progressed to the extraction of more complicated features such as Fisher vectors,[74] before the advent of deep learning made these steps obsolete. In fact, the interpretability of deep networks is at the semantic level of images, not features, for example, to answer the question of whether the network takes the tail of a dog into account when predicting its race. This is not easily possible in radiomics, where a visualization of the important areas of a tumor lesion would not help a radiologist understand what the network is doing. Furthermore, in current machine learning, a model is accepted if it generalizes well, not necessarily if the model is interpretable. A similar strategy may be viable for radiomics, where the applicability of models is validated on large datasets.

In conclusion, radiomics currently faces substantial challenges related to imaging and statistical reproducibility that severely impact interpretability and clinical applicability. These problems are difficult to mitigate because imaging standardization is largely impractical and statistical variability is inherent in high-dimensional datasets. As a result, the potential for clinical integration remains uncertain and questionable. A shift toward rigorous data and code sharing practices and the development of large, representative datasets would be required to partially address these challenges.

Footnotes

Conflict of interest disclosure

The author declared no conflicts of interest.

References

1. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-446. [CrossRef]
2. McCague C, Ramlee S, Reinius M, et al. Introduction to radiomics for a clinical audience. *Clin Radiol*. 2023;78(2):83-98. [CrossRef]
3. Haneberg AG, Pierre K, Winter-Reinhold E, et al. Introduction to radiomics and artificial intelligence: a primer for radiologists. *Semin Roentgenol*. 2023;58(2):152-157. [CrossRef]
4. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006. [CrossRef]
5. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;30(9):1234-1248. [CrossRef]
6. Harlow CA, Dwyer SJ, Lodwick G. On radiographic image analysis. In: *Digital Picture Analysis*. Springer; 1976:65-150. [CrossRef]
7. Freeborough PA, Fox NC. MR image texture analysis applied to the diagnosis and tracking of Alzheimer's disease. *IEEE Trans Med Imaging*. 1998;17(3):475-478. [CrossRef]
8. Crombé A, Bertolo F, Fadli D, et al. Distinct patterns of the natural evolution of soft tissue sarcomas on pre-treatment MRIs captured with delta-radiomics correlate with gene expression profiles. *Eur Radiol*. 2023;33(2):1205-1218. [CrossRef]
9. Qi Y, Zhao T, Han M. The application of radiomics in predicting gene mutations in cancer. *Eur Radiol*. 2022;32(6):4014-4024. [CrossRef]
10. Xia TY, Zhou ZH, Meng XP, et al. Predicting microvascular invasion in hepatocellular carcinoma using CT-based radiomics model. *Radiology*. 2023;307(4):e222729. [CrossRef]
11. Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö. Radiomics with artificial intelligence: a practical guide for beginners. *Diagn Interv Radiol*. 2019;25(6):485-495. [CrossRef]
12. Guiot J, Vaidyanathan A, Deprez L, et al. A review in radiomics: making personalized medicine a reality via routine imaging. *Med Res Rev*. 2022;42(1):426-440. [CrossRef]
13. Koçak B, Cuocolo R, Santos DPD, Stanzione A, Ugga L. Must-have qualities of clinical research on artificial intelligence and machine learning. *Balkan Med J*. 2023;40(1):3-12. [CrossRef]
14. Teno JM. Garbage in, garbage out-words of caution on big data and machine learning in medical practice. *JAMA Health Forum*. 2023;4(2):e230397. [CrossRef]
15. Zhao B. Understanding sources of variation to improve the reproducibility of radiomics. *Front Oncol*. 2021;11:633176. [CrossRef]
16. Ibrahim A, Barufaldi B, Refaee T, et al. MaasPenn radiomics reproducibility score: a novel quantitative measure for evaluating the reproducibility of CT-based handcrafted radiomic features. *Cancers (Basel)*. 2022;14(7):1599. [CrossRef]
17. Schwier M, Griethuysen J van, Vangel MG, et al. Repeatability of multiparametric prostate MRI radiomics features. *Sci Rep*. 2019;9(1):9441. [CrossRef]
18. Emaminejad N, Wahi-Anwar MW, Kim GHJ, Hsu W, Brown M, McNitt-Gray M. Reproducibility of lung nodule radiomic features: multivariable and univariable investigations that account for interactions between CT acquisition and reconstruction parameters. *Med Phys*. 2021;48:2906-2919. [CrossRef]
19. Marcadent S, Hofmeister J, Preti MG, Martin SP, Van De Ville D, Montet X. Generative adversarial networks improve the reproducibility and discriminative power of radiomic features. *Radiol Artif Intell*. 2020;2(3):e190035. [CrossRef]
20. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT radiomics. *Radiology*. 2019;291(1):53-59. [CrossRef]
21. Granzier RWY, Verbakel NMH, Ibrahim A, et al. MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability. *Sci Rep*. 2020;10(1):14163. [CrossRef]
22. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328-338. [CrossRef]
23. Bettinelli A, Marturano F, Avanzo M, et al. A novel benchmarking approach to assess the agreement among radiomic tools. *Radiology*. 2022;303(3):533-541. [CrossRef]
24. Whybra P, Zwanenburg A, Andrearczyk V, et al. The image biomarker standardization initiative: standardized convolutional filters for reproducible radiomics and enhanced clinical insights. *Radiology*. 2024;310(2):e231319. [CrossRef]
25. Demircioğlu A. The effect of preprocessing filters on predictive performance in radiomics. *Eur Radiol Exp*. 2022;6(1):40. [CrossRef]
26. Pappu V, Pardalos PM. High-dimensional data classification. In: Aleskerov F, Goldengorin B, Pardalos PM, eds. *Clusters, Orders, and Trees: Methods and Applications: In Honor of Boris Mirkin's 70th Birthday*. Springer New York; 2014:119-150. [CrossRef]
27. Demircioğlu A. Benchmarking feature selection methods in radiomics. *Invest Radiol*. 2022;57(7):433-443. [CrossRef]
28. Ge G, Siddique A, Zhang J. Inconsistent CT NSCLC radiomics associated with feature selection methods, predictive models and related factors. *Phys Med Biol*. 2023;68(12):125004. [CrossRef]
29. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B Methodol*. 1996;58(1):267-288. [CrossRef]
30. Peng H, Ding C. Minimum redundancy and maximum relevance feature selection and recent advances in cancer classification: 8. [CrossRef]
31. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw*. 2010;36(11):1-13. [CrossRef]
32. Krawczuk J, Łukaszuk T. The feature selection bias problem in relation to high-dimensional gene data. *Artif Intell Med*. 2016;66:63-71. [CrossRef]
33. Demircioğlu A. Measuring the bias of incorrect application of feature selection when using

cross-validation in radiomics. *Insights Imaging*. 2021;12(1):172. [CrossRef]

34. Doran SJ, Kumar S, Orton M, et al. "Real-world" radiomics from multi-vendor MRI: an original retrospective study on the prediction of nodal status and disease survival in breast cancer, as an exemplar to promote discussion of the wider issues. *Cancer Imaging*. 2021;21(1):37. [CrossRef]
35. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. [CrossRef]
36. Lo SB, Lou SA, Lin JS, Freedman MT, Chien MV, Mun SK. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Trans Med Imaging*. 1995;14(4):711-718. [CrossRef]
37. Sahiner B, Chan HP, Petrick N, et al. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE Trans Med Imaging*. 1996;15(5):598-610. [CrossRef]
38. Blazis SP, Dickerscheid DBM, Linsen PVM, Martins Jarnalo CO. Effect of CT reconstruction settings on the performance of a deep learning based lung nodule CAD system. *Eur J Radiol*. 2021;136:109526. [CrossRef]
39. Qin Z, Yu F, Liu C, Chen X. How convolutional neural network see the world - a survey of convolutional neural network visualization methods. Published online May 31, 2018. Accessed June 26, 2024. [CrossRef]
40. Banerjee S, Mitra S, Masulli F, Rovetta S. Glioma classification using deep radiomics. *SN Comput Sci*. 2020;1(4):209. [CrossRef]
41. Zhang X, Zhang G, Qiu X, et al. Radiomics under 2D regions, 3D regions, and peritumoral regions reveal tumor heterogeneity in non-small cell lung cancer: a multicenter study. *Radiol Med*. 2023;128(9):1079-1092. [CrossRef]
42. Müller D, Kramer F. MISCnn: a framework for medical image segmentation with convolutional neural networks and deep learning. *BMC Med Imaging*. 2021;21(1):12. [CrossRef]
43. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. [CrossRef]
44. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009:248-255. [CrossRef]
45. Demircioğlu A. Deep features from pretrained networks do not outperform hand-crafted features in radiomics. *Diagnostics (Basel)*. 2023;13(20):3266. [CrossRef]
46. Li Z, Wang Y, Yu J, Guo Y, Cao W. Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. *Sci Rep*. 2017;7(1):5467. [CrossRef]
47. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? Published online July 18, 2022. [CrossRef]
48. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med Phys*. 2017;44(10):5162-5171. [CrossRef]
49. Han W, Qin L, Bay C, et al. Deep transfer learning and radiomics feature prediction of survival of patients with high-grade gliomas. *Am J Neuroradiol*. 2020;41(1):40-48. [CrossRef]
50. Demircioğlu A. Are deep models in radiomics performing better than generic models? A systematic review. *Eur Radiol Exp*. 2023;7(1):11. [CrossRef]
51. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci*. 2001;16(3):199-231. [CrossRef]
52. Bellman R. Dynamic programming. *Science*. 1966;153(3731):34-37. [CrossRef]
53. Demircioğlu A. The effect of feature normalization methods in radiomics. *Insights Imaging*. 2024;15(1):2. [CrossRef]
54. Demircioğlu A. Evaluation of the dependence of radiomic features on the machine learning model. *Insights Imaging*. 2022;13(1):28. [CrossRef]
55. Tohidinezhad F, Di Perri D, Zegers CML, et al. Prediction models for radiation-induced neurocognitive decline in adult patients with primary or secondary brain tumors: a systematic review. *Front Psychol*. 2022;13:853472. [CrossRef]
56. Costa G, Cavinato L, Fiz F, et al. Mapping tumor heterogeneity via local entropy assessment: making biomarkers visible. *J Digit Imaging*. 2023;36(3):1038-1048. [CrossRef]
57. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: the need for safeguards. *Radiother Oncol*. 2019;130:2-9. [CrossRef]
58. Manikis GC, Ioannidis GS, Siakallis L, et al. Multicenter DSC-MRI-based radiomics predict IDH mutation in gliomas. *Cancers*. 2021;13(16):3965. [CrossRef]
59. Amparore E, Perotti A, Bajardi P. To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods. *PeerJ Comput Sci*. 2021;7:e479. [CrossRef]
60. Lou Y, Caruana R, Gehrke J, Hooker G. Accurate intelligible models with pairwise interactions. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. Association for Computing Machinery; 2013:623-631. [CrossRef]
61. Cho HH, Lee HY, Kim E, et al. Radiomics-guided deep neural networks stratify lung adenocarcinoma prognosis from CT scans. *Commun Biol*. 2021;4(1):1286. [CrossRef]
62. Breznau N, Rinke EM, Wuttke A, et al. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proc Natl Acad Sci*. 2022;119(44):e2203150119. [CrossRef]
63. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. *Sci Transl Med*. 2021;13(586):eabb1655. [CrossRef]
64. Yoon JH, Strand F, Baltzer PAT, et al. Standalone AI for breast cancer detection at screening digital mammography and digital breast tomosynthesis: a systematic review and meta-analysis. *Radiology*. 2023;307(5):e222639. [CrossRef]
65. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *IEEE Conference on Computer Vision and Pattern Recognition*. 2017:3462-3471. [CrossRef]
66. Kocak B, Akinci D'Antonoli T, Mercaldo N, et al. METhodological RadiomICs score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII. *Insights Imaging*. 2024;15(1):8. [CrossRef]
67. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-762. [CrossRef]
68. Moskowitz CS, Welch ML, Jacobs MA, Kurland BF, Simpson AL. Radiomic analysis: study design, statistical analysis, and other bias mitigation strategies. *Radiology*. 2022;304(2):265-273. [CrossRef]
69. Bao J, Qiao X, Song Y, et al. Prediction of clinically significant prostate cancer using radiomics models in real-world clinical practice: a retrospective multicenter study. *Insights Imaging*. 2024;15(1):68. [CrossRef]
70. Kocak B, Borgheresi A, Ponsiglione A, et al. Explanation and elaboration with examples for CLEAR (CLEAR-E3): an EuSoMII radiomics auditing group initiative. *Eur Radiol Exp*. 2024;8:72. [CrossRef]
71. Ioannidis JP. Why most published research findings are false. *PLOS Med*. 2005;2(8):e124. [CrossRef]
72. Serra-Garcia M, Gneezy U. Nonreplicable publications are cited more than replicable ones. *Sci Adv*. 2021;7(21):eabd1705. [CrossRef]
73. Kocak B, Bulut E, Bayrak ON, et al. NEgatiVE results in radiomics research (NEVER): a meta-research study of publication bias in leading radiology journals. *Eur J Radiol*. 2023;163:110830. [CrossRef]
74. Sánchez J, Perronnin F, Mensink T, Verbeek J. Image classification with the Fisher vector: theory and practice. *Int J Comput Vis*. 2013;105(3):222-245. [CrossRef]
75. Choyke P, Turkbey B, Pinto P, Merino M, Wood B. Data from PROSTATE-MRI. The cancer imaging archive. *Cancer Imaging Arch TCIA*. Published online 2016. [CrossRef]