

SYSTEMATIC REVIEW

Open Access



Are deep models in radiomics performing better than generic models? A systematic review

Aydin Demircioğlu*

Abstract

Background Application of radiomics proceeds by extracting and analysing imaging features based on generic morphological, textural, and statistical features defined by formulas. Recently, deep learning methods were applied. It is unclear whether deep models (DMs) can outperform generic models (GMs).

Methods We identified publications on PubMed and Embase to determine differences between DMs and GMs in terms of receiver operating area under the curve (AUC).

Results Of 1,229 records (between 2017 and 2021), 69 studies were included, 61 (88%) on tumours, 68 (99%) retrospective, and 39 (56%) single centre; 30 (43%) used an internal validation cohort; and 18 (26%) applied cross-validation. Studies with independent internal cohort had a median training sample of 196 (range 41–1,455); those with cross-validation had only 133 (43–1,426). Median size of validation cohorts was 73 (18–535) for internal and 94 (18–388) for external. Considering the internal validation, in 74% (49/66), the DMs performed better than the GMs, vice versa in 20% (13/66); no difference in 6% (4/66); and median difference in AUC 0.045. On the external validation, DMs were better in 65% (13/20), GMs in 20% (4/20) cases; no difference in 3 (15%); and median difference in AUC 0.025. On internal validation, fused models outperformed GMs and DMs in 72% (20/28), while they were worse in 14% (4/28) and equal in 14% (4/28); median gain in AUC was +0.02. On external validation, fused model performed better in 63% (5/8), worse in 25% (2/8), and equal in 13% (1/8); median gain in AUC was +0.025.

Conclusions Overall, DMs outperformed GMs but in 26% of the studies, DMs did not outperform GMs.

Keywords Artificial intelligence, Deep learning, Machine learning, Radiology, Radiomics

Key Points

- Deep learning (DL) models outperform generic models often but only in 3 out of 4 studies.
- Fused models can improve over the generic and DL models.
- Data leakage, model selection and optimisation, and publication bias could affect the comparison between generic and DL models.
- It is worthwhile to explore both modelling strategies in practice.

Background

The application of machine learning (ML) to radiological imaging is a fairly old idea and can be traced back at least to the 1970s [1]. There are many benefits to such an approach. First, it can be noninvasively applied to various tasks like diagnosis, classification, and prognosis [2–4].

*Correspondence:
Aydin Demircioğlu
aydin.demircioglu@uk-essen.de
Institute of Diagnostic and Interventional Radiology and Neuroradiology,
University Hospital Essen, Hufelandstraße 55, 45147 Essen, Germany

Second, because imaging potentially contains more information than humans can process, ML can exploit it systematically, which could lead to superhuman performance. Finally, it allows for automation of time-consuming tasks, saving human resources for other essential duties.

While this idea was explored in several studies in the 1990s and 2000s under the heading of texture analysis [5–7], it prominently resurfaced only in 2012, when it was coined “radiomics” in a seminal paper by Lambin et al. [8]. A classic ML pipeline applied to radiomics consists of several steps (Fig. 1) [9]. It was shown that radiomics can lead to accurate models [10–12].

Despite its benefits, there are also disadvantages to this approach. A key problem is the segmentation of pathologic findings, which is often tedious work. Even though some (semi)automated solutions exist [13], they are not always ready for diagnostic purposes. A second disadvantage is the set of features that are used to characterise the pathologies quantitatively; these are defined generically by the use of explicit formulas and are often derived from morphological, statistical, and textural properties [14]. We call such features “generic” since they are not specific to the problem at hand and could thus be unable to capture all information present in the data. Hence, models based on generic features, *i.e.*, “generic models” (GMs), could perform suboptimally. In contrast, models could learn features directly from datasets during training without requiring explicit formulas. This approach, however, requires different methods.

Deep learning (DL), which is a subarea of ML based on neural networks, could be able to solve the

disadvantages of generic models. While the main idea dates back to the first ML concepts in the 1950s [15], and networks already have been applied to radiological data in the 1990s [16], only recently, new techniques and increased computation power allowed these networks to solve many interesting problems that were previously thought to be hard.

Compared to generic modelling, the DL pipeline involves fewer steps (Fig. 1), as deep networks can learn directly from images without the explicit need for any segmentations. Equally important, they can learn predictive features independently during training, bypassing the need for explicit feature definitions. We will refer to models that use features learned implicitly from data by DL as “deep models” (DMs). Because deep models are adapted to data, it is reasonable to expect them to yield better results than GMs.

Yet, there are drawbacks to deep modelling. Since the networks are not given previously defined features, they usually need more data to find predictive patterns. However, larger sample sizes are often unavailable in a radiomic context. In addition, the reproducibility and generalizability of deep networks are unclear since they are known to be sensitive to the initial weights and might behave erratically [17–19]. Both might render any advantage of deep modelling against generic modelling void.

Nonetheless, since DL methods have been applied successfully in many fields, they are generally considered to be superior to generic modelling in radiomics, even though large-scale experiments that analyse this question in-depth are currently missing. Indeed, some radiomics studies using DL report higher predictive performance

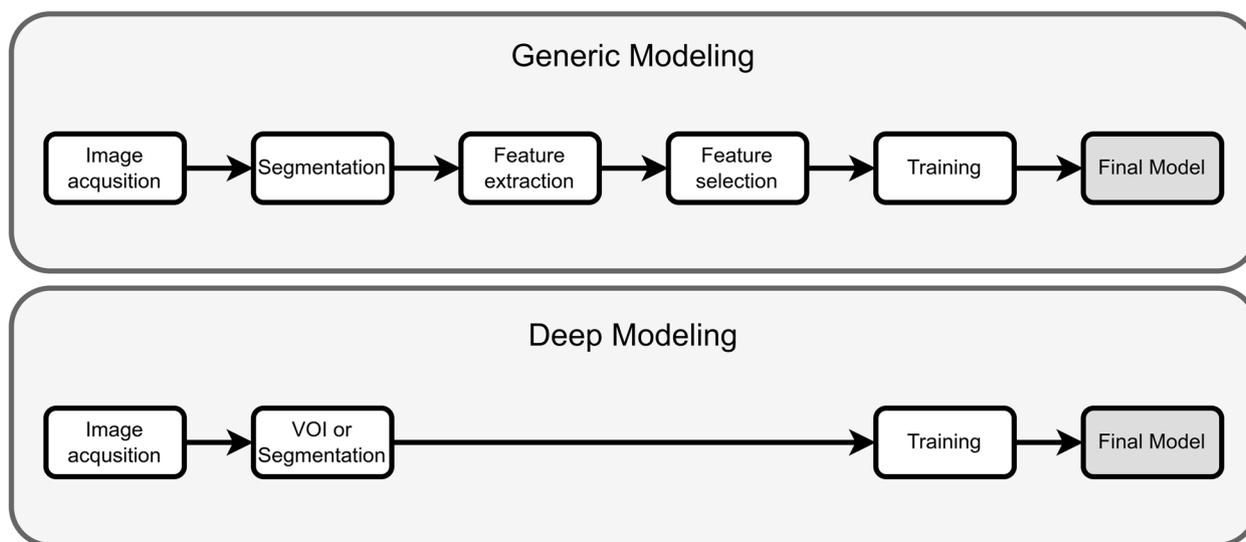


Fig. 1 Generic and deep modelling applied to radiomics

than generic modelling [20, 21], but other studies report no improvements [22, 23].

This observation raises the question of whether deep modelling is truly superior. Therefore, in this review, we examined studies that directly compare deep and GMs to determine whether a difference in predictive performance can be seen. We also discuss the influence of a few modelling decisions and potential biases for differences in performance between the two models.

Modelling strategies

Generic and deep modelling have many processing steps in common (Fig. 1). The main difference relates to the generation of features. While in generic modelling standardised features are extracted from the imaging data, deep modelling can employ a wide range of network architectures to find optimal features.

Generic modelling

The overall pipeline of generic modelling is rather standardised (Fig. 1) [3, 24]. Starting from image acquisition, a key step is to segment the pathologies. This is necessary to focus the computation of the features on the relevant area of the imaging. Features are then extracted from the volumes and are used to train a classifier.

In all these steps, certain choices have to be performed. For example, the imaging data is first discretised to avoid that the extracted features depend too much on the inherent noise [25, 26]. This discretisation proceeds either by binning the data into bins of fixed widths or a predefined number of bins, and it is unclear which of the two approaches works better for a given dataset. From these discretised images, many features, like maximum, minimum, mean, and variance, are computed using explicit formulas, which characterise the image in a specific way. However, since it is not known beforehand which features will be predictive, many irrelevant and redundant features will be present. Feature selection methods are applied to remove these, and the remaining features are then processed by a classifier [27].

While the choice of the feature selection method and classifier is central to obtaining a high-performing model, in our study, we took the GM as a baseline and therefore only considered decision choices regarding the deep modelling.

Deep modelling

Deep modelling can be more complex than generic modelling, as neural networks can be built with many different topologies and architectures. A description of deep networks is beyond the scope of this review; details can be found in the literature [28–30]. In a nutshell, a deep network can be described as a model with multiple layers

of neurons connected by weights. These weights are used to transform input data into output data; for example, a scan depicting a tumour could be transformed into a prediction of its malignancy. The weights, therefore, are central since they determine the network's output. Training of a network can then be understood as a process to optimise the weights so that input data is transformed to the corresponding label.

However, deep networks are parameterised by a vast number of weights, numbering in the millions. Thus, a sufficiently large number of training samples is required for successful training, making them unsuitable per se in areas such as radiomics, where only limited sample sizes are available. Pretraining is a commonly used trick to get around this problem, where the network is trained on data from another domain. The hope is that by pretraining, the weights will be in a near-optimal state so that for successful training of the problem at hand, fewer samples are necessary.

Pretraining, however, cannot be directly applied to radiomic data since many pretrained networks were trained on photographs and can, therefore, only process two-dimensional (2D) data, while radiomic data is often three-dimensional (3D). A solution would be to process the radiomic data slice by slice, but in this approach, the spatial context is lost, and the network's performance will be suboptimal. On the other hand, employing a 3D network is also difficult because of the low sample sizes; pretrained 3D networks are also currently unavailable, further deepening the problem. Therefore, a critical choice in developing a DM is whether the network should be 2D or 3D and whether pretraining should be used.

Networks can also be trained in an end-to-end fashion or used as feature extractors. In the end-to-end case, the network is trained and used as a whole. In contrast, when used as a feature extractor, features are extracted at an intermediate layer of the network and then processed using classical machine learning methods. The advantage of this approach is that other techniques can be used for classification, possibly improving the overall performance.

Fused models (FMs)

Models that fuse generic and deep models are of particular interest, as they should be able to harness the advantages of both modelling strategies and lead to yet higher predictive performance since fusing works similarly to a small ensemble [31]. The fusion can take place on multiple levels; the most basic approach is to take the average of the output of both models. If the network is used as a feature extractor, another approach would be to merge generic and deep features and apply feature selection and classification methods to this merged feature set [32].

Alternatively, the generic features can be added directly to the network, so it can utilise them during training [33]. More complicated fusion methods are also possible [34, 35], but because generic features are rather fixed, the options mainly affect the neural network's architecture.

Fusing does not come without disadvantages; the FM depends on the GM, which in turn usually depends on fine segmentation. Therefore, the key advantage of deep modelling is lost in FMs. In addition, since the FM will have more hyperparameters, the risk of overfitting is higher [36].

Literature review

We conducted a literature search to find studies comparing the two modelling approaches to gather evidence on their relative performance.

Search protocol

We identified publications published before 2022 by querying PubMed and Embase databases using the keywords “radiomics” and either “deep learning”, “deep neural”, or “deep network” (Supplementary material 1).

Study selection

Abstracts of the publications were first screened; all studies which were not original research were removed. Studies were excluded that either did not report on a binary outcome or were not using 3D data. Studies were not considered eligible if as follows: (a) no GM was trained or reported; (b) no DM was trained or reported; (c) only a FM was trained or reported; (d) deep learning was not used for modelling; (e) no area under the curve (AUC) was reported; and (f) the validation scheme was unclear. In addition, AUCs from segmentation or survival tasks were not included in our study since these AUCs differ; for example, a segmentation is a low-dimensional problem, which has different statistical properties than the high-dimensional problems that usually occur in radiomic studies with a binary outcome.

Study questions

The main question of this study was *whether models based on deep modelling perform better than GMs in terms of predictive performance, measured by AUC*. In addition, we aimed to answer the following questions: (a) Do DMs perform better on external data than GMs? (b) Do FMs perform better than either the GMs or DMs? (c) Do 3D network architectures outperform 2D networks? (d) Does pretraining help to improve predictive performance? (e) Does the use of deep networks as feature extractors increase the gain in AUC over end-to-end learning? These questions concern only deep modelling since we consider GMs as the baseline.

Data extraction

We listed the sample size and the validation scheme used for each study. We then extracted the predictive performance of the generic, deep, and FMs in the internal and external cohorts. Internal data refers here to data collected at the same hospital or centre; external data is those gathered at a different hospital or centre. Predictive performance was extracted as mean and 95%-confidence intervals or converted from standard deviation where possible [37]. The difference between the AUC of the generic, deep, and FMs was then computed and graphically displayed separately for models tested on internal and external data.

Statistics

Descriptive statistics results were reported as median. Statistics were computed using Python 3.10.4.

Results

Literature research

Of 1229 records, 69 studies were included in the analysis (Fig. 2, Supplementary material 2). An overview of all included studies is presented in Table 1. Publication dates ranged from 2017 to 2022, with most studies conducted in 2021 (Fig. 3); no relevant study was found before 2017, whereas a few studies were available in 2021 ahead of final publication and therefore had a publication date of 2022.

Study characteristics

Except for one study [38], all were retrospective in nature. Studies varied greatly in their sample size (Table 2). In the studies that used an independent internal cohort, the median sample size of the training cohorts was 196 (range 41–1,455). In contrast, in the studies that employed cross-validation only, without an independent internal cohort, the training sample size was even smaller (133, range 43–1,426). The validation cohorts were also smaller than the training cohorts; the internal cohorts had a median size of 73 samples (range 18–535), and the external cohorts had 94 samples (range 18–388).

Nearly all studies related to tumours (88%). More studies used computed tomography than magnetic resonance imaging (50% versus 40%) and were conducted on only one site (56%); accordingly, most studies either used an internal validation cohort (43%) or applied cross-validation (26%).

Predictive performance

Comparing the performance of generic and deep modelling on the internal validation sets, in 74% (49/66), the

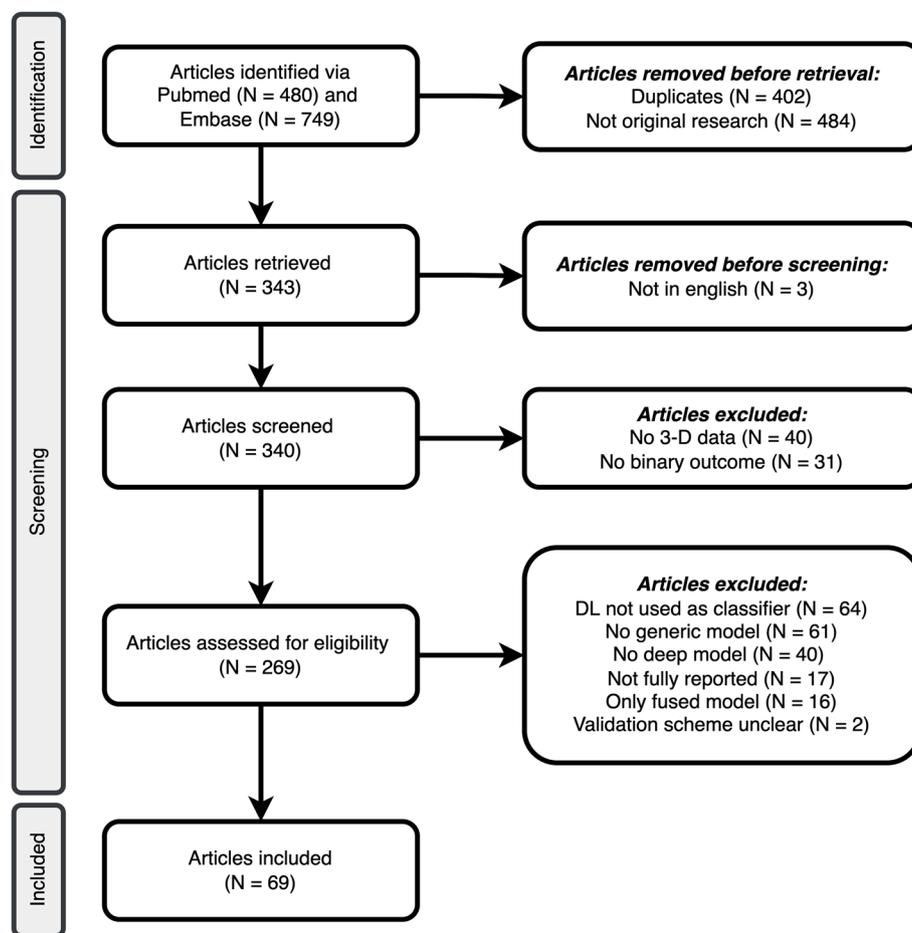


Fig. 2 Inclusion and exclusion flowchart

DMs performed better than the GMs and vice versa in 20% (13/66) of the cases (Fig. 4). In 6% (4/66), there was no difference. The median difference in AUC was 0.045.

On the external validation sets, the DMs were better in 65% (13/20) of the cases and the GMs in 20% (4/20) cases (Fig. 5). In three cases (15%), no difference was seen. The median difference in AUC was 0.025.

A similar picture emerged when considering the FMs. On the internal validation sets, the FMs outperformed the better of GMs and the DMs in 72% (20/28), while it was worse in 14% (4/28) and equal in 14% (4/28) (Fig. 6). The median performance gain in AUC was +0.02. On the external validation sets, the FM performed better in 63% of the cases (5/8), worse in 25% (2/8), and equal in a single case (13%) (Fig. 7). The median gain in AUC was +0.025.

Characteristics of deep modelling

Nearly all DMs were based on convolutional neural networks (CNN) (96%, 66/69), with only three exceptions that used either a capsule network [39], a generative

adversarial network [40], or a sparse autoencoder [41]. Note that two studies employed a U-Net, which is generative CNN [42, 43]. Most DMs used 2D architectures (58%, 40/69), while pretraining was performed less frequently (45%, 31/69). In 78% (54/69) of all studies, the network was trained using the data; in the remaining 22% (15/69), no training was performed. The network predictions were used directly in 55% (38/69) of the cases, while in 45% (31/69), the network was used as a feature extractor.

Regarding the characteristics of the deep networks, 2D networks performed better than 3D networks when comparing both to the GMs: the median performance gain in AUC in the internal cohorts for 2D networks was, on average, +0.05 and for 3D networks +0.02 (Table 3). A slightly higher gain could be seen in the external cohorts (+0.08 *versus* +0.0). Furthermore, pretraining yielded a higher performance gain (internal cohorts: median AUC +0.07 *versus* +0.02; external cohorts: +0.09 *versus* +0.01). Using the network in an end-to-end fashion or as a feature extractor did not

Table 1 Overview of the characteristics of all included studies

Study author	Year	Pathology	Validation	Multicentre	Modality	DOI
Cha	2017	Bladder cancer	Internal cohort	●	CT	10.1038/s41598-017-09315-w
Li	2017	Brain cancer	LOO-CV	●	MRI	10.1038/s41598-017-05848-2
Oakden-Rayner	2017	Longevity	6-fold CV	●	CT	10.1038/s41598-017-01931-w
Hosny	2018	Lung cancer	External cohort	●	CT	10.1371/journal.pmed.1002711
Chen	2019	Head and neck cancer	Internal cohort	●	PET/CT	10.1088/1361-6560/ab083a
Diamant	2019	Head and neck cancer	External cohort	●	CT	10.1038/s41598-019-39206-1
Shinde	2019	Parkinson's disease	Internal cohort	●	MRI	10.1016/j.micl.2019.101748
Zhao	2019	Lung cancer	External cohort	●	CT	10.1002/cam4.2233
Zhu	2019	Brain cancer	External cohort	●	MRI	10.1016/j.ejrad.2019.04.022
Feng	2020	Lung cancer	External cohort	●	CT	10.1007/s00330-020-07024-z
Fu	2020	Rectal cancer	4-fold CV	●	MRI	10.1088/1361-6560/ab7970
He	2020	Colorectal cancer	Internal cohort	●	CT	10.1186/s12880-020-00457-4
Hua	2020	Cervical cancer	Internal cohort	●	MRI	10.1016/j.bspe.2020.101869
Li	2020	COVID-19	Internal cohort	●	CT	10.1109/JBHI.2020.3036722
Sun	2020	Gastric cancer	Internal cohort	●	CT	10.1016/j.ejrad.2020.109277
Wang+Li	2020	Lung cancer	Internal cohort	●	CT	10.21037/tlcr-20-370
Whitney	2020	Breast cancer	Internal cohort	●	MRI	10.1109/jproc.2019.2950187
Xia	2020	Lung cancer	External cohort	●	CT	10.3389/fonc.2020.00418
Yan	2020	Cervical cancer	Internal cohort	●	MRI	10.1259/bjr.20200358
Zhao	2020	Lung cancer	5-fold CV	●	MRI	10.1016/j.jlucan.2020.04.014
Zhou	2020	Breast cancer	10-fold CV	●	MRI	10.1002/jmri.26981
Ziegelhayer	2020	Pancreatic cancer	2-fold CV	●	CT	10.3390/jcm9124013
Aompong	2021	Lung cancer	10-fold CV	●	CT	10.1109/EMBC46164.2021.9630703
Astaraki	2021	Lung cancer	Internal cohort	●	CT	10.3389/fonc.2021.737368
Avanzo	2021	Lung cancer	10-fold CV	●	CT+3D-BED	10.1002/mp.15178
Bo	2021	Brain cancer	Internal cohort	●	MRI	10.3389/fmed.2021.748144
Caballo	2021	Breast cancer	Internal cohort	●	CT	10.1117/1.JMI.8.2.024501
Castillo	2021	Prostate cancer	External cohort	●	MRI	10.3390/cancers14010012
Chen+Dohopolski	2021	Head and neck cancer	5-fold CV	●	PET/CT	10.1016/j.jirobp.2021.02.004
Chen+Lin	2021	Brain cancer	Internal cohort	●	MRI	10.3389/fonc.2021.734433
Gao	2021	Liver cancer	Internal cohort	●	MRI	10.1016/j.mri.2021.06.018
Guo	2021	Lung cancer	Internal cohort	●	CT	10.1016/j.acra.2020.06.010
Han	2021	Lung cancer	Internal cohort	●	PET/CT	10.1007/s00259-020-04771-5
Hou	2021	Prostate cancer	External cohort	●	MRI	10.1016/j.ebiom.2021.103395
Hu+Gong	2021	Lung cancer	Internal cohort	●	CT	10.1088/1361-6560/abe735
Hu+Xie	2021	Lung cancer	External cohort	●	CT	10.1016/j.radonc.2020.09.014
Jiang+ Cao	2021	Liver cancer	Internal cohort	●	CT	10.1007/s00432-020-03366-9
Jiang+Ren	2021	Lung cancer	Internal cohort	●	MRI	10.1002/jmri.27579
Kang	2021	Sarcoma	External cohort	●	CT	10.3389/fonc.2021.750875
Li	2021	Lung cancer	Internal cohort	●	CT	10.1016/j.compmedimag.2021.102009
Lin	2021	Lung cancer	5-fold CV	●	CT	10.1016/j.elle.2021.02.004
Liu	2021	Lung cancer	4-fold CV	●	CT	10.21037/qims-20-734
Marentakis	2021	Lung cancer	2-fold CV	●	CT	10.1007/s11517-020-02302-w
Naglah	2021	Thyroid cancer	LOO-CV	●	MRI	10.3390/s21113878
Ning	2021	Brain cancer	External cohort	●	MRI	10.21037/atm-20-4076
Peng	2021	Liver cancer	External cohort	●	CT	10.3389/fonc.2021.730282
Shao	2021	Placenta invasion	External cohort	●	MRI	10.3934/mbe.2021310
Song+Ding	2021	Lung cancer	External cohort	●	CT	10.1002/mp.15307
Song+Wang+Luo	2021	Pancreatic cancer	External cohort	●	CT	10.21037/atm-21-25
Song+Wang+Wang	2021	Liver cancer	Internal cohort	●	MRI	10.1007/s00432-021-03617-3
Tian	2021	Liver cancer	5-fold CV	●	MRI	10.3390/diagnostics11101875
Wan	2021	Gastric cancer	Internal cohort	●	CT	10.1002/mp.15199
Wang+Dong	2021	COVID-19	External cohort	●	CT	10.1109/JBHI.2021.3076086
Wang+Wang	2021	COVID-19	External cohort	●	CT	10.1007/s00259-020-05075-4
Wang+Yuan	2021	Lung cancer	5-fold CV	●	CT	10.1002/mp.15293
Wei	2021	Liver cancer	4-fold CV	●	CT	10.1002/mp.14563
Wu	2021	Gastrointestinal cancer	External cohort	●	CT	10.1002/mp.14767
Xia	2021	Brain cancer	5-fold CV	●	MRI	10.1002/jmri.27592
Xiao+He	2021	Parkinson's disease	7-fold CV	●	MRI	10.3389/fmins.2021.760975
Xiao+Sun	2021	Breast cancer	Internal cohort	●	DBT	10.1186/s12938-021-00908-1
Xuan	2021	Placenta invasion	Internal cohort	●	MRI	10.1186/s12938-021-00893-5
Zhang	2021	Head and neck cancer	External cohort	●	MRI	10.1002/jmri.27308
Bertelli	2022	Prostate cancer	Internal cohort	●	MRI	10.3389/fonc.2021.802964
Jiang	2022	Breast cancer	External cohort	●	DBT	10.1002/mp.15392
Li	2022	Brain cancer	Internal cohort	●	MRI	10.1007/s00330-021-08237-6
Liu	2022	Sarcoma	External cohort	●	MRI	10.1007/s00330-021-08221-0
Xu	2022	Renal cancer	Internal cohort	●	MRI	10.1002/jmri.27900
Yang+Wang	2022	Head and neck cancer	Internal cohort	●	CT	10.1007/s00066-021-01874-2
Yang+Zhou	2022	Liver cancer	Internal cohort	●	CT	10.1016/j.ejso.2021.11.120

If multiple outcomes were analysed in a study, only the outcome with the highest area under the curve was reported. Note that Li (2020), Astaraki (2021), Caballo (2021), Chen + Lin (2021), Li (2021), Lin (2021), and Xuan (2021) used data from external sites; however, the data were merged and randomly split before modeling. Therefore, the results were considered to be internally validated, not externally. ● yes, ● no, 3D-BED Three-dimensional biologically effective dose, COVID-19 Coronavirus disease 2019, CV Cross-validation, CT Computed tomography, DBT Digital breast tomosynthesis, LOO-CV Leave-on-out-cross-validation, MRI Magnetic resonance imaging, PET Positron emission tomography

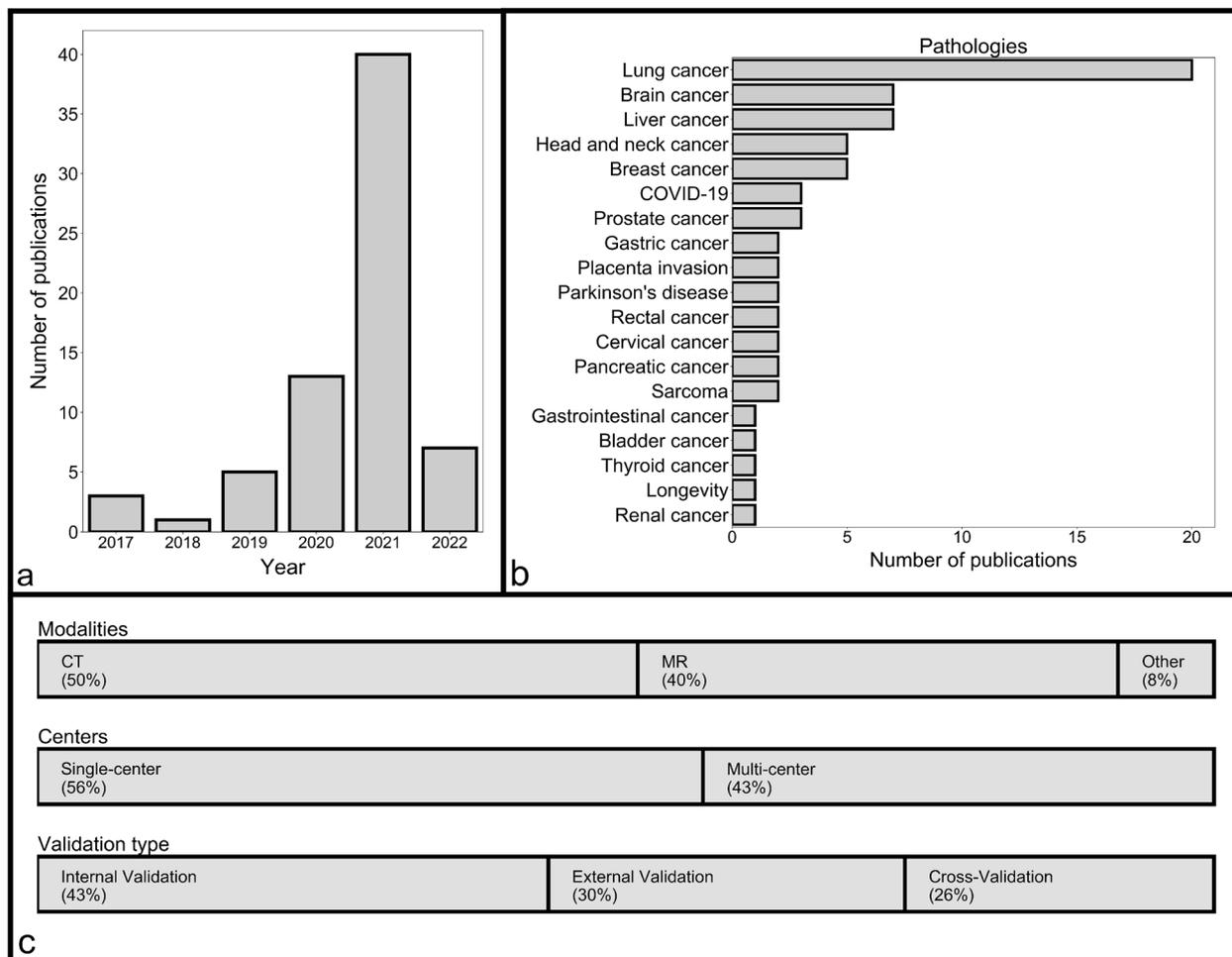


Fig. 3 Characteristics of the included studies

make a clear difference in the internal cohorts (median AUC +0.05 for both). The situation was different in the external cohorts (+0.02 *versus* 0.09).

Discussion

Generic and DL models are currently in use in radiomics, but the difference in predictive performance has not yet been analysed across studies. Therefore, we reviewed studies that used both modelling strategies and identified 69 studies that provided a direct comparison.

Predictive performance

Overall, there was a significant advantage of deep over generic modelling, as evident by an increase in median AUC of +0.045 in the internal cohorts. However, in 26% of the studies, no increase was visible. It is unclear whether this depended on the data or if the modelling was not performed as well as it could have

been. A lower difference was found in external cohorts (AUC +0.025), indicating that DMs perform on external data at least as good as GMs. Fusing the two modelling approaches had similar gains (AUC +0.02 and +0.025). Since the overall number of studies with a FM was smaller, the effect must be read cautiously.

Characteristics of DL modelling

As expected, network architectures derived directly from CNN were used nearly exclusively, and other architectures were vastly underexplored. Therefore, we considered three deep modelling choices that are relevant for CNNs: the dimensionality of the network, the use of pre-trained weights, and end-to-end training. It turned out that on average, 2D networks performed better than 3D networks when compared to the GMs (AUC +0.05 *versus* +0.03); nonetheless, the median sample sizes of the training sets were higher for 3D networks (224 *versus* 166 samples), showing that most studies preferred 3D

Table 2 Overview of the predictive performance reported in the included studies

Study authors	Year	Multicentre	#CV	#Training	#Test (internal)	AUC (deep, internal)	AUC (generic, internal)	AUC difference (deep versus generic), internal	AUC (fused), internal	AUC difference (fused versus better), internal	#Test (external)	AUC (deep, external)	AUC (generic, external)	AUC difference (deep versus generic), external	AUC (fused), external	AUC difference (fused versus better), external	Network dimension	Network architecture	Network pretrained?	Network trained?	Network as feature extractor?
Cha	2017	●	–	82	41	0.77	0.73	-0.04	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
Li	2017	●	110	–	–	0.86	0.92	0.06	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
Oakden-Rayner	2017	●	48	–	–	0.65	0.68	0.03	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Hosny	2018	●	–	290	–	–	–	–	–	–	101	0.58	0.71	0.13	–	–	3D	CNN	●	●	●
Chen	2019	●	–	41	18	0.89	0.95	0.06	0.95	0.0	–	–	–	–	–	–	3D	CNN	●	●	●
Diamant	2019	●	–	194	106	0.86	0.88	0.02	0.92	0.04	–	–	–	–	–	–	2D	CNN	●	●	●
Shinde	2019	●	–	55	25	0.54	0.91	0.37	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
Zhao	2019	●	–	464	115	0.64	0.76	0.12	–	–	37	0.69	0.75	0.06	–	–	3D	CNN	●	●	●
Zhu	2019	●	–	99	82	0.68	0.81	0.13	0.82	0.01	–	–	–	–	–	–	2D	CNN	●	●	●
Feng	2020	●	–	218	140	0.7	0.8	0.1	–	–	192	0.65	0.77	0.12	–	–	2D	CNN	●	●	●
Fu	2020	●	43	–	–	0.64	0.73	0.09	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
He	2020	●	–	117	40	0.82	0.93	0.11	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
Hua	2020	●	–	111	67	0.73	0.69	-0.04	0.78	0.05	–	–	–	–	–	–	2D	CNN	●	●	●
Li	2020	●	–	174	43	0.84	0.79	-0.05	0.86	0.02	–	–	–	–	–	–	3D	CNN	●	●	●
Sun	2020	●	–	252	176	0.8	0.88	0.08	–	–	144	–	–	–	–	–	2D	CNN	●	●	●
Wang+Li	2020	●	–	753	133	0.94	0.89	-0.05	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Whitney	2020	●	–	1455	535	0.89	0.89	0.0	0.91	0.02	–	–	–	–	–	–	2D	CNN	●	●	●
Xia	2020	●	–	229	–	–	–	–	–	–	94	0.87	0.83	-0.04	0.9	0.03	3D	CNN	●	●	●
Yan	2020	●	–	100	90	0.72	0.72	0.0	0.82	0.1	–	–	–	–	–	–	2D	CNN	●	●	●
Zhao	2020	●	501	–	–	0.87	0.88	0.01	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Zhou	2020	●	133	–	–	0.91	0.99	0.08	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
Ziegelmeier	2020	●	86	–	–	0.8	0.9	0.1	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
Aonpong	2021	●	211	–	–	0.68	0.69	0.01	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
Astaraki	2021	●	–	1266	506	0.92	0.94	0.02	0.94	0.0	–	–	–	–	–	–	2D	CNN	●	●	●
Avanzo	2021	●	76	–	–	0.77	0.81	0.04	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Bo	2021	●	–	131	57	0.75	0.71	-0.04	0.85	0.1	–	–	–	–	–	–	2D	CNN	●	●	●
Caballo	2021	●	–	202	82	0.79	0.84	0.05	0.88	0.04	–	–	–	–	–	–	2D	CNN	●	●	●
Castillo	2021	●	271	–	–	0.83	0.89	0.06	–	–	195	0.91	0.73	-0.18	–	–	3D	CNN	●	●	●
Chen+Dohopolski	2021	●	129	–	–	0.95	0.96	0.01	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Chen+Lin	2021	●	–	170	74	0.83	0.84	0.01	0.91	0.07	–	–	–	–	–	–	3D	CNN	●	●	●
Gao	2021	●	–	168	57	0.65	0.77	0.12	0.83	0.06	–	–	–	–	–	–	2D	CNN	●	●	●
Guo	2021	●	–	782	138	0.79	0.84	0.05	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Han	2021	●	–	1136	283	0.86	0.9	0.04	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
Hou	2021	●	–	280	71	0.83	0.84	0.01	0.85	0.01	50	0.78	0.78	0.0	0.77	-0.01	2D	CNN	●	●	●
Hu+Gong	2021	●	–	365	64	0.65	0.62	-0.03	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Hu+Xie	2021	●	161	–	–	0.82	0.9	0.08	0.8	-0.1	70	0.72	0.8	0.08	–	–	2D	CNN	●	●	●
Jiang+Cao	2021	●	–	324	81	0.87	0.91	0.04	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Jiang+Ren	2021	●	–	77	20	0.77	0.64	-0.13	0.77	0.0	–	–	–	–	–	–	2D	CNN	●	●	●
Kang	2021	●	–	241	104	0.84	0.9	0.06	–	–	388	0.78	0.81	0.03	–	–	3D	CNN	●	●	●
Li	2021	●	–	237	58	0.75	0.83	0.08	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Lin	2021	●	915	–	–	0.87	0.83	-0.04	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Liu	2021	●	126	–	–	0.84	0.85	0.01	–	–	–	–	–	–	–	–	3D	CapsNet	●	●	●
Marentakis	2021	●	102	–	–	0.67	0.78	0.11	0.69	-0.09	–	–	–	–	–	–	2D	CNN	●	●	●
Naglah	2021	●	49	–	–	0.85	0.85	0.0	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Ning	2021	●	–	186	47	0.92	0.91	-0.01	0.94	0.02	334	0.85	0.84	-0.01	0.88	0.03	3D	CNN	●	●	●
Peng	2021	●	139	–	–	0.97	0.98	0.01	–	–	171	0.96	0.97	0.01	–	–	2D	CNN	●	●	●
Shao	2021	●	–	78	–	–	–	–	–	–	34	0.92	0.92	0.0	–	–	3D	CNN	●	●	●
Song+Ding	2021	●	–	528	–	0.72	0.82	0.1	–	–	137	0.68	0.79	0.11	–	–	2D	CNN	●	●	●
Song+Wang+Luo	2021	●	56	–	–	0.74	0.8	0.06	–	–	18	0.56	0.77	0.21	–	–	2D	CNN	●	●	●
Song+Wang+Wang	2021	●	–	461	140	0.73	0.92	0.19	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Tian	2021	●	103	–	–	0.79	0.85	0.06	0.9	0.05	–	–	–	–	–	–	3D	CNN	●	●	●
Wan	2021	●	–	96	41	0.76	0.77	0.01	0.83	0.06	–	–	–	–	–	–	2D	SAE	●	●	●
Wang+Dong	2021	●	–	271	68	0.86	0.76	-0.1	0.88	0.02	61	0.6	0.75	0.15	0.86	0.11	3D	CNN	●	●	●
Wang+Wang	2021	●	–	184	18	0.96	0.92	-0.04	0.97	0.01	32	0.85	0.86	0.01	0.86	0.0	2D	GAN	●	●	●
Wang+Yuan	2021	●	1426	–	–	0.91	0.9	-0.01	0.94	0.03	200	0.9	0.88	-0.02	0.92	0.02	3D	CNN	●	●	●
Wei	2021	●	192	–	–	0.6	0.82	0.22	0.83	0.01	–	–	–	–	–	–	2D	CNN	●	●	●
Wu	2021	●	333	–	–	0.92	0.92	0.0	0.89	-0.03	183	0.89	0.91	0.02	–	–	2D	CNN	●	●	●
Xia	2021	●	289	–	–	0.91	0.96	0.05	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
Xiao+He	2021	●	140	–	–	0.88	0.95	0.07	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Xiao+Sun	2021	●	–	211	25	0.84	0.85	0.01	–	–	–	–	–	–	–	–	3D	CNN	●	●	●
Xuan	2021	●	–	189	163	0.8	0.88	0.08	0.9	0.02	–	–	–	–	–	–	3D	CNN	●	●	●
Zhang	2021	●	–	169	19	0.74	0.58	-0.16	–	–	45	0.68	0.68	0.0	–	–	2D	CNN	●	●	●
Bertelli	2022	●	–	93	19	0.8	0.85	0.05	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
Jiang	2022	●	–	147	73	0.77	0.79	0.02	0.82	0.03	46	0.58	0.67	0.09	0.79	0.12	2D	CNN	●	●	●
Li	2022	●	–	780	236	0.84	0.89	0.05	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
Liu	2022	●	–	70	–	0.85	0.92	0.07	0.92	0.0	43	0.56	0.78	0.22	0.76	-0.02	2D	CNN	●	●	●
Xu	2022	●	–	173	44	0.83	0.92	0.09	–	–	–	–	–	–	–	–	2D	CNN	●	●	●
Yang+Wang	2022	●	–	208	89	0.66	0.81	0.15	0.69	-0.12	–	–	–	–	–	–	2D	CNN	●	●	●
Yang+Zhou	2022	●	–	198	85	0.74	0.88	0.14	–	–	–	–	–	–	–	–	2D	CNN	●	●	●

If only a cross-validation was performed without an independent internal validation cohort, the number of training samples across all folds is reported the #CV column. Otherwise, if internal validation cohort was available, then the number of training and validation samples are reported in the #training and #test column. Note that Li (2020), Astaraki (2021), Caballo (2021), Chen + Lin (2021), Li (2021), Lin (2021), and Xuan (2021) used data from external sites; however, the data were merged and randomly split before modeling. Therefore, the results were considered to be internally validated, not externally. Note also that Hu + Gong (2021) and Song + Wang + Luo (2021) use a U-Net, which is a generative network. ● yes, ● no, 2D Two-dimensional, 3D Three-dimensional, CapsNet Capsule neural network, CNN Convolutional neural network, GAN Generative adversarial network, SAE Sparse autoencoder

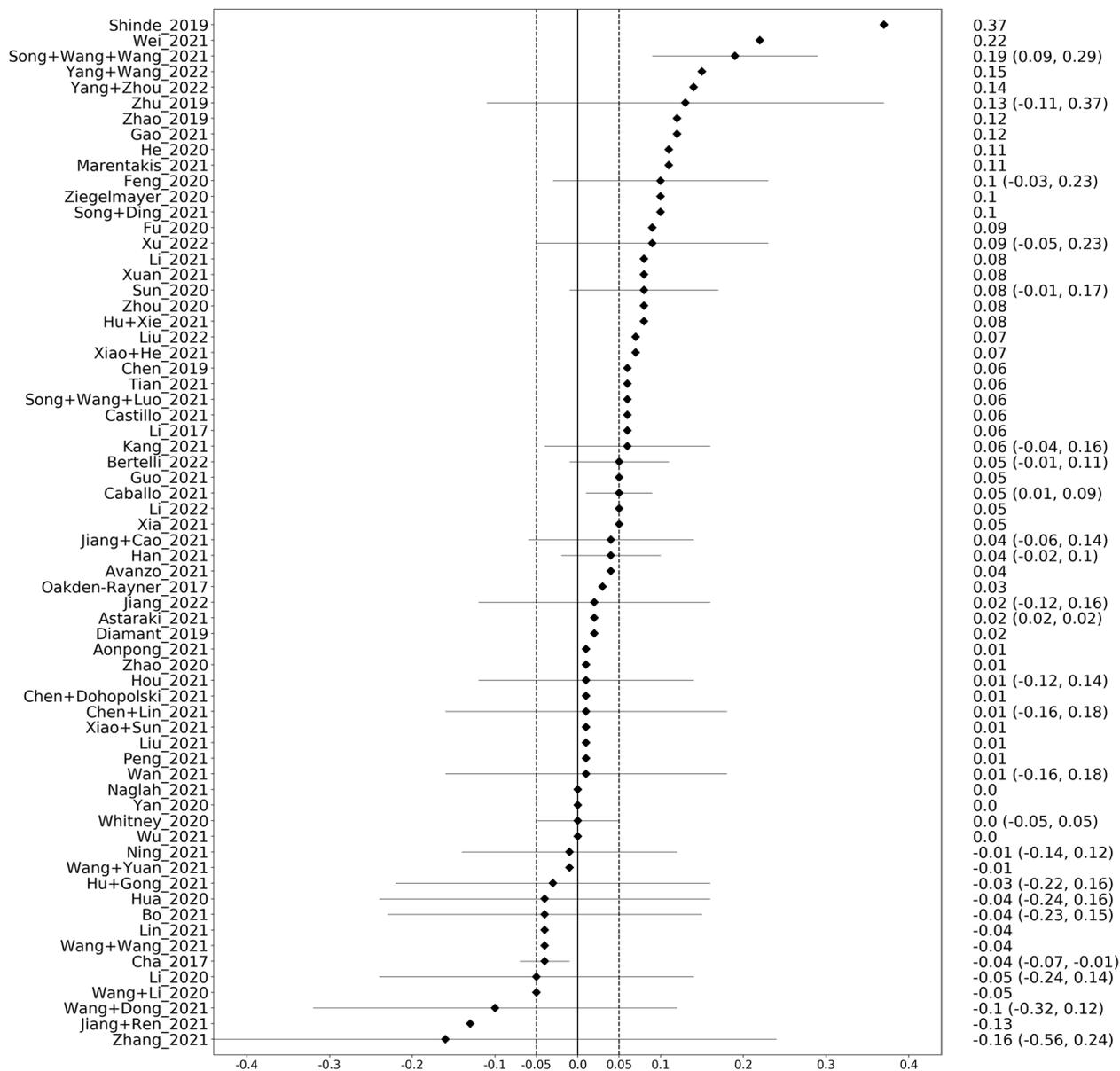


Fig. 4 Graphical display of the performance differences between the generic and deep models on the internal validation sets. On the right, the difference in area under the curve together with the 95% confidence interval is given. A positive difference means that the deep model performed better than the generic model

networks when sample sizes were higher. Pretraining did yield a higher performance gain (AUC 0.07 *versus* 0.02). Finally, using the network in an end-to-end fashion or as a feature extractor did not make a difference (AUC +0.05 *versus* +0.05).

These observations must be taken with some caution since they only reflect an average tendency. For example, pretraining did not yield superior results in some studies [22, 44, 45]; similarly, 3D networks can perform better [46].

Sample size

The data sets used were relatively small on average, as reflected in the median sample sizes of the training cohorts ($n=196$). However, the sample sizes of the test cohorts were even smaller ($n=73$ and $n=94$). Thus, caution should be exercised when using such small cohorts to demonstrate that one modelling approach is statistically better than another. Because a rule of thumb derived from simple statistical distributions requires at least 30

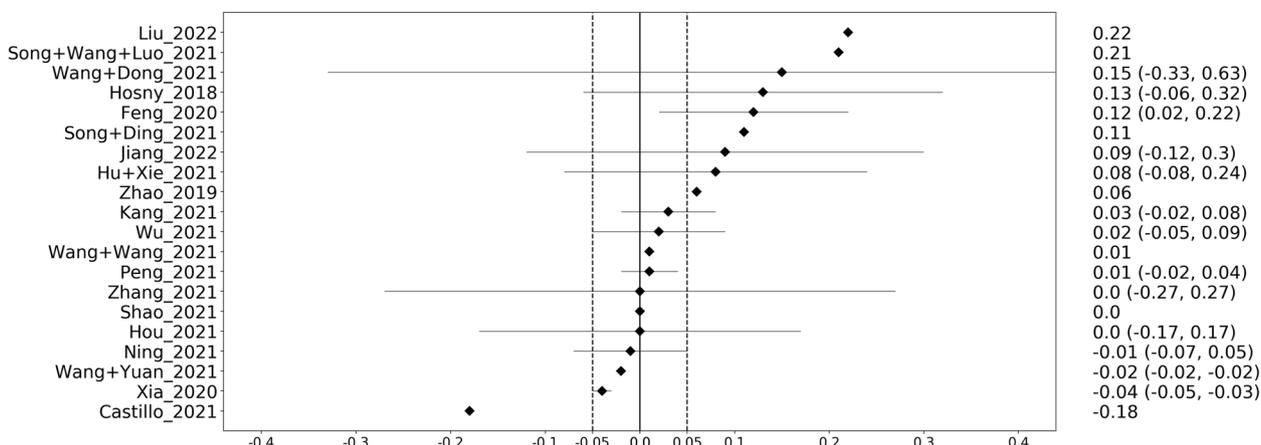


Fig. 5 Graphical display of the performance differences between the generic and deep models on the external validation sets. On the right, the difference in area under the curve together with the 95% confidence interval is given. A positive difference means that the deep model performed better than the generic model

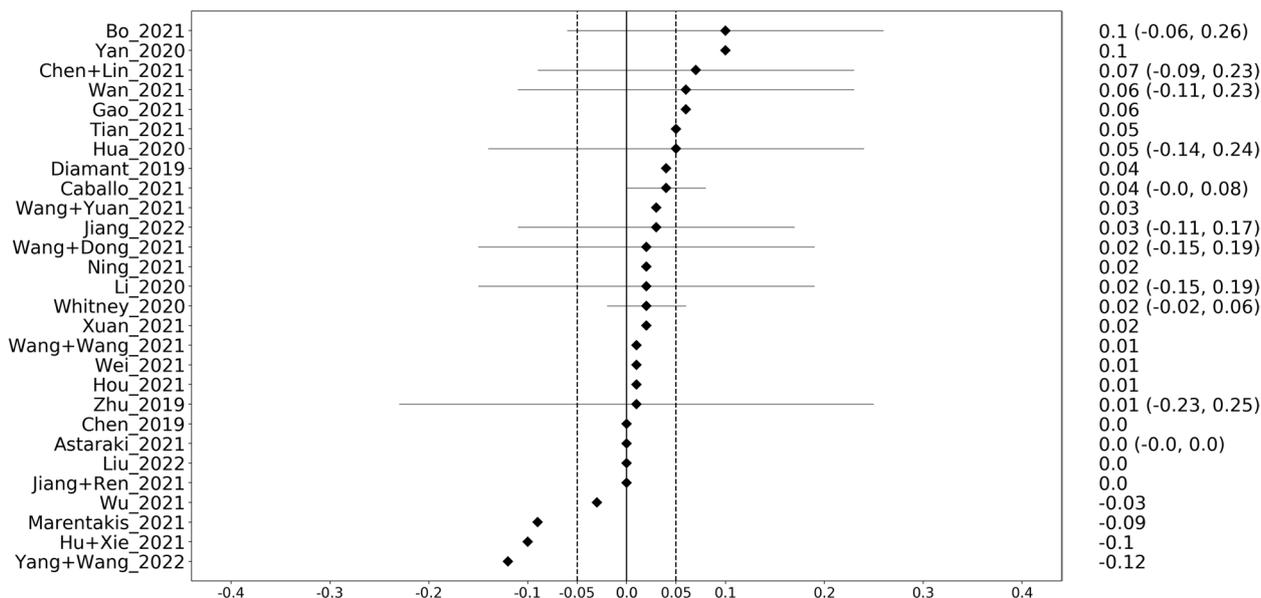


Fig. 6 Graphical display of the performance differences between the fused and the better of generic and deep models on the internal validation sets. On the right, the difference in area under the curve together with the 95% confidence interval is given. A positive difference means that the fused model performed better than the deep and the generic models

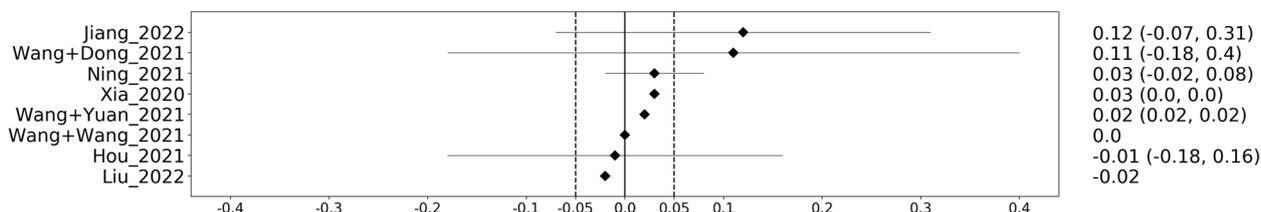


Fig. 7 Graphical display of the performance differences between the fused and the better of generic and deep models on the external validation sets. On the right, the difference in area under the curve together with the 95% confidence interval is given. A positive difference means that the fused model performed better than the deep and the generic models

Table 3 Overview of the influence of network characteristics on the predictive performance relative to generic modelling

Network characteristic		Internal validation cohorts				External validation cohorts			
		Median gain in AUC	Better	Equal	Worse	Median gain in AUC	Better	Equal	Worse
Dimension	Two-dimensional	+0.05	78% (31/40)	8% (3/40)	15% (6/40)	+0.08	82% (9/11)	0% (0/11)	18% (2/11)
	Three-dimensional	+0.02	69% (18/26)	4% (1/26)	27% (7/26)	+0.00	44% (4/9)	11% (1/9)	44% (4/9)
Weights	Pretrained	+0.07	86% (24/28)	7% (2/28)	7% (2/28)	+0.09	67% (6/9)	22% (2/9)	11% (1/9)
	Trained from scratch	+0.02	66% (25/38)	5% (2/38)	29% (11/38)	+0.01	64% (7/11)	9% (1/11)	27% (3/11)
Approach	End-to-end	+0.05	72% (26/36)	8% (3/36)	19% (7/36)	+0.02	60% (6/10)	10% (1/10)	30% (3/10)
	Feature extractor	+0.05	77% (23/30)	3% (1/30)	20% (6/30)	+0.09	70% (7/10)	20% (2/10)	10% (1/10)

The median gain in area under the curve (AUC) was calculated as the difference in performance from the generic models across all studies that used a network with the corresponding feature. Similarly, the “better”, “equal”, and “worse” columns denote the number of studies that reported better, equal, or worse AUC of the deep model (with the corresponding feature) compared with the generic model

samples per group to establish a statistical difference [47, 48], the sample sizes of the studies appear to be too small.

Validation and generalisability

Many of the included studies used an internal test cohort instead of applying cross-validation. However, testing on a single split can be unreliable [49], and since cross-validation can be understood as applying systematically repeated splits, it should be preferred. Most of the studies also did not test their model on external data, since setting up a large-scale multicentre study has a high organisational cost. Their generalisability, measured as performance on external data sets, is therefore unclear. In addition, since nearly all studies were retrospective in nature, their clinical applicability [10] was not tested.

Sources of bias

Several sources of bias exist which can impede a fair comparison. These include biases in modelling and study quality, interpretation, and publication.

Data leakage

In both modelling strategies, data leakage can occur; for generic modelling, a few studies seemed to apply the feature selection to all data before applying cross-validation and do not test their model on a fully independent cohort [50]; however, this can lead to a strong positive bias [51]. The same error can occur for DMs if they are used as feature extractors. For DMs, it is also conceivable that some studies misused the validation cohort for testing during training, which would also lead to a positive bias. These problems cannot be detected from the studies, adding complexity to a direct comparison.

Interpretability

In the literature, generic features are thought to be more interpretable by a human reader than deep features [14].

Therefore, a potential trade-off occurs when using deep features, since a gain in predictive performance would come with a loss in interpretability. This trade-off cannot be directly quantified since it is unclear how to measure “interpretable”. While some generic features like volume and mean intensity have a clear meaning, other features like “Wavelet-HLH_glrIm_RunVariance” are inherently obscure. Nevertheless, the characteristics of the generic features utilised by the model can potentially be used in further experiments to demonstrate a biological correlation, which is not directly possible for deep features. It must be noted, however, that statistically equivalent models may select quite different features [52].

Bias in modelling

For both modelling strategies, there is a certain bias concerning the used methods. In general, it is unknown a priori which methods will perform best; therefore, it is best practice to test multiple methods. For example, feature selection is crucial in higher dimensions, and different choices can lead to models with vastly different performance [27, 53, 54]. Yet, some of the included studies only consider a single feature selection and classification method for generic modelling [55]; if this method is not adequate for the data, it can lead to an underperforming model and would introduce another bias. Other studies extract only a few features [56] which can also result in a potential loss in performance [57]. DMs can also suffer from such a bias since nearly all studies used a CNN architecture; it is conceivable that other network architectures could perform better [58].

In both modelling strategies, hyperparameters, which are parameters that are not learned during training, need to be selected. This is critical for many models; for example, a network can only perform well if the learning rate is chosen properly. But since tuning hyperparameters is

computationally expensive, hyperparameters are often left at their default values, which can lead to degraded performance. This can be problematic if tuning is only applied to one of both models; perceived improvements would be spurious since the comparisons could not be regarded as fair.

Such unfair comparisons might happen more often than expected because many papers aim to show that DMs can yield higher performances than GMs. Therefore, it might happen unintentionally that most efforts will be put into developing the DM, whereas less effort is put into the GM. This is especially true if the DM initially performs worse than expected. In this case, deep modelling might be continued until a better-performing model is found.

Study quality

A well-known problem in radiomics concerns the quality of studies, which can lead to a lack of reproducibility. Although guidelines exist [59], and a quality score specific to radiomic papers has been introduced [9], many papers still need to adhere to this guideline [60]. Thus, the overall quality of studies may be another bias factor that needs to be considered.

Publication bias

On top, publishing negative results is still not common. Thus, if studies are undertaken with the hope that the DM improves over the GM, but fail to show this, the results might not be reported. Furthermore, a widely underperforming DM might also be not reported since there is a certain risk that training was performed incorrectly; this is true for the GMs to a far lesser extent since the training is less complex and there are more user-friendly ML frameworks [61, 62].

Recommendations

Given these results, we recommend not limiting oneself to a GM or a DM but computing both. Care should be taken that the GM is not neglected, but that the full range of methods and parameters is tested. Deep modelling with pretrained 2D networks based on CNN architectures is advisable, although, if permissible, custom 3D network architectures should also be tested. Finally, it also seems to make sense to test a FM as it can improve the predictive performance even further.

Limitations

Our study has a few limitations. First, a direct comparison between GMs and DMs is influenced by many factors, for example the preprocessing and harmonisation

of the images and the choice of segmentations, that is, whether they include tissue beyond the pathology. However, to keep things simple, we only considered a few factors for deep modelling and none for generic modelling, since we regarded these as a baseline. We also only considered AUCs, the most often used measurement, though other measures like sensitivity and specificity are often equally important. In addition, several papers report on multiple methods and outcomes. In these cases, we selected the model or outcome with the highest AUC (regardless whether it was obtained in the training or test cohort). These choices could have potentially introduced a bias into our study.

Conclusion and future directions

In this review, evidence has been found that deep modelling can outperform generic modelling. However, since this is not always the case, both generic and deep modelling should be considered in radiomics. Even though our results showed that DL outperforms generic modelling by some margin, the comparison was only indirect. A large benchmark study involving several datasets would lead to a better understanding of the modelling strategies and yield more precise recommendations. This would include studies on the reproducibility of deep features since these were only performed for generic features yet [63, 64]. DL also has many more applications in radiomics that need to be explored in detail. For example, it can be used as an image-to-image transformer [65] and for automated segmentations [66]; these possibilities are orthogonal to both modelling strategies and could improve both.

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
AUC	Area under the curve
CNN	Convolutional neural network
DL	Deep learning
DM	Deep model
FM	Fused model
GM	Generic model
ML	Machine learning

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s41747-023-00325-0>.

Additional file 1.

Additional file 2.

Authors' contributions

AD is the only author of this study. AD conceptualised the review, performed all searches, collected all data, evaluated the data, and wrote the manuscript. The author read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. The author declares that no funds, grants, or other support was received during the preparation of this manuscript.

Availability of data and materials

All data generated or analysed during this study are included in this published article (and its supplementary information files).

Declarations**Ethics approval and consent to participate**

This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to publication

Not applicable.

Competing interests

The author declares no competing interests.

Received: 30 September 2022 Accepted: 13 January 2023

Published online: 15 March 2023

References

- Harlow CA, Dwyer SJ, Lodwick G (1976) On radiographic image analysis. In: Rosenfeld A (ed), *Digital Picture Analysis*. Springer, Heidelberg, pp 65–150
- Yip SSF, Aerts HJWL (2016) Applications and limitations of radiomics. *Phys Med Biol* 61:R150–R166. <https://doi.org/10.1088/0031-9155/61/13/R150>
- Scapicchio C, Gabelloni M, Barucci A, Cioni D, Saba L, Neri E (2021) A deep look into radiomics. *Radiol Med* 126:1296–1311. <https://doi.org/10.1007/s11547-021-01389-x>
- Guiot J, Vaidyanathan A, Deprez L et al (2022) A review in radiomics: making personalized medicine a reality via routine imaging. *Med Res Rev* 42:426–440. <https://doi.org/10.1002/med.21846>
- Schad LR, Blüml S, Zuna I (1993) IX. MR tissue characterization of intracranial tumors by means of texture analysis. *Magn Reson Imaging* 11:889–896. [https://doi.org/10.1016/0730-725X\(93\)90206-S](https://doi.org/10.1016/0730-725X(93)90206-S)
- Gibbs P, Turnbull LW (2003) Textural analysis of contrast-enhanced MR images of the breast. *Magn Reson Med* 50:92–98. <https://doi.org/10.1002/mrm.10496>
- Kovalev VA, Kruggel F, Gertz H-J, von Cramon DY (2001) Three-dimensional texture analysis of MRI brain datasets. *IEEE Trans Med Imaging* 20:424–433. <https://doi.org/10.1109/42.925295>
- Lambin P, Rios-Velazquez E, Leijenaar R et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* 48:441–446. <https://doi.org/10.1016/j.ejca.2011.11.036>
- Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 14:749–762. <https://doi.org/10.1038/nrclinonc.2017.141>
- Satake H, Ishigaki S, Ito R, Naganawa S (2022) Radiomics in breast MRI: current progress toward clinical application in the era of artificial intelligence. *Radiol Med* 127:39–56. <https://doi.org/10.1007/s11547-021-01423-y>
- Kang D, Park JE, Kim Y-H et al (2018) Diffusion radiomics as a diagnostic model for atypical manifestation of primary central nervous system lymphoma: development and multicenter external validation. *Neuro Oncol* 20:1251–1261. <https://doi.org/10.1093/neuonc/nyo021>
- Davenport T, Kalakota R (2019) The potential for artificial intelligence in healthcare. *Future Healthc J* 6:94–98. <https://doi.org/10.7861/futurehosp.6-2-94>
- Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 18:203–211. <https://doi.org/10.1038/s41592-020-01008-z>
- Abbasian Ardakani A, Bureau NJ, Ciaccio EJ, Acharya UR (2022) Interpretation of radiomics features—a pictorial review. *Comput Methods Programs Biomed* 215:106609. <https://doi.org/10.1016/j.cmpb.2021.106609>
- Fradkov AL (2020) Early history of machine learning IFAC-Pap 53:1385–1390. <https://doi.org/10.1016/j.ifacol.2020.12.1888>
- Lo SB, Lou SA, Lin JS, Freedman MT, Chien MV, Mun SK (1995) Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Trans Med Imaging* 14:711–718. <https://doi.org/10.1109/42.476112>
- Lee Y, Oh S-H, Kim MW (1991) The effect of initial weights on premature saturation in back-propagation learning. In: *IJCNN-91-Seattle International Joint Conference on Neural Networks*. pp 765–770 vol.1
- Huang G, Li Y, Pleiss G, et al (2017) Snapshot ensembles: Train 1, get M for free. *ArXiv170400109 Cs*
- Dauphin YN, Pascanu R, Gulcehre C, et al (2014) Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: *Proc. Advances in Neural Information Processing Systems* 27, pp 2933–2941
- Ziegelmayr S, Reischl S, Harder F, Makowski M, Braren R, Gawlitza J (2022) Feature robustness and diagnostic capabilities of convolutional neural networks against radiomics features in computed tomography imaging. *Invest Radiol* 57:171–177. <https://doi.org/10.1097/RLI.0000000000000827>
- Song D, Wang Y, Wang W et al (2021) Using deep learning to predict microvascular invasion in hepatocellular carcinoma based on dynamic contrast-enhanced MRI combined with clinical parameters. *J Cancer Res Clin Oncol* 147:3757–3767. <https://doi.org/10.1007/s00432-021-03617-3>
- Whitney HM, Li H, Ji Y, Liu P, Giger ML (2020) Comparison of breast MRI tumor classification using human-engineered radiomics, transfer learning from deep convolutional neural networks, and fusion methods. *Proc IEEE Inst Electr Electron Eng* 108:163–177. <https://doi.org/10.1109/JPROC.2019.2950187>
- Wang S, Dong D, Li L et al (2021) A deep learning radiomics model to identify poor outcome in COVID-19 patients with underlying health conditions: a multicenter study. *IEEE J Biomed Health Inform* 25:2353–2362. <https://doi.org/10.1109/JBHI.2021.3076086>
- Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö (2019) Radiomics with artificial intelligence: a practical guide for beginners. *Diagn Interv Radiol* 25:485–495. <https://doi.org/10.5152/dir.2019.19321>
- Duron L, Balvay D, Vande Perre S, et al (2019) Gray-level discretization impacts reproducible MRI radiomics texture features. *PLoS One* 14:e0213459. <https://doi.org/10.1371/journal.pone.0213459>
- 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>
- Demircioğlu A (2022) Benchmarking feature selection methods in radiomics. *Invest Radiol* <https://doi.org/10.1097/RLI.0000000000000855>
- Parekh VS, Jacobs MA (2019) Deep learning and radiomics in precision medicine. *Expert Rev Precis Med Drug Dev* 4:59–72. <https://doi.org/10.1080/23808993.2019.1585805>
- Maier A, Syben C, Lasser T, Riess C (2019) A gentle introduction to deep learning in medical image processing. *Z Med Phys* 29:86–101. <https://doi.org/10.1016/j.zemedi.2018.12.003>
- Goodfellow I, Bengio Y, Courville A (2016) *Deep learning*. MIT Press, Cambridge
- Tulyakov S, Jaeger S, Govindaraju V, Doermann D (2008) Review of classifier combination methods. In: *Marinai S, Fujisawa H (eds) Machine learning in document analysis and recognition*. Springer, Berlin, Heidelberg, pp 361–386
- Bo L, Zhang Z, Jiang Z, et al (2021) Differentiation of brain abscess from cystic glioma using conventional MRI based on deep transfer learning features and hand-crafted radiomics features. *Front Med (Lausanne)* 8:748144. <https://doi.org/10.3389/fmed.2021.748144>
- Cheng H-T, Ispir M, Anil R, et al (2016) Wide & deep learning for recommender systems. In: *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems - DLRS 2016*. ACM Press, Boston, MA, USA, pp 7–10
- Caballo M, Pangallo DR, Mann RM, Sechopoulos I (2020) Deep learning-based segmentation of breast masses in dedicated breast CT imaging: radiomic feature stability between radiologists and artificial intelligence.

- Comput Biol Med 118:103629. <https://doi.org/10.1016/j.combiomed.2020.103629>
35. Gao F, Qiao K, Yan B et al (2021) Hybrid network with difference degree and attention mechanism combined with radiomics (H-DARnet) for MVI prediction in HCC. *Magn Reson Imaging* 83:27–40. <https://doi.org/10.1016/j.mri.2021.06.018>
 36. Hosseini M, Powell M, Collins J et al (2020) I tried a bunch of things: the dangers of unexpected overfitting in classification of brain data. *Neurosci Biobehav Rev* 119:456–467. <https://doi.org/10.1016/j.neubiorev.2020.09.036>
 37. Higgins JPT, Thomas J, Chandler J et al (2019) *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons
 38. Sun R-J, Fang M-J, Tang L, et al (2020) CT-based deep learning radiomics analysis for evaluation of serosa invasion in advanced gastric cancer. *Eur J Radiol* 132:109277. <https://doi.org/10.1016/j.ejrad.2020.109277>
 39. Liu H, Jiao Z, Han W, Jing B (2021) Identifying the histologic subtypes of non-small cell lung cancer with computed tomography imaging: a comparative study of capsule net, convolutional neural network, and radiomics. *Quant Imaging Med Surg* 11:2756–2765. <https://doi.org/10.21037/qims-20-734>
 40. Wang H, Wang L, Lee EH et al (2021) Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures. *Eur J Nucl Med Mol Imaging* 48:1478–1486. <https://doi.org/10.1007/s00259-020-05075-4>
 41. Wan Y, Yang P, Xu L et al (2021) Radiomics analysis combining unsupervised learning and handcrafted features: a multiple-disease study. *Med Phys* 48:7003–7015. <https://doi.org/10.1002/mp.15199>
 42. Hu X, Gong J, Zhou W, et al (2021) Computer-aided diagnosis of ground glass pulmonary nodule by fusing deep learning and radiomics features. *Phys Med Biol* 66:065015. <https://doi.org/10.1088/1361-6560/abe735>
 43. Song C, Wang M, Luo Y, et al (2021) Predicting the recurrence risk of pancreatic neuroendocrine neoplasms after radical resection using deep learning radiomics with preoperative computed tomography images. *Ann Transl Med* 9:833–833. <https://doi.org/10.21037/atm-21-25>
 44. Diamant A, Chatterjee A, Vallières M, et al (2019) Deep learning in head & neck cancer outcome prediction. *Sci Rep* 9. <https://doi.org/10.1038/s41598-019-39206-1>
 45. Marentakis P, Karaiskos P, Kouloulis V et al (2021) Lung cancer histology classification from CT images based on radiomics and deep learning models. *Med Biol Eng Comput* 59:215–226. <https://doi.org/10.1007/s11517-020-02302-w>
 46. Chen L, Zhou Z, Sher D, et al (2019) Combining many-objective radiomics and 3D convolutional neural network through evidential reasoning to predict lymph node metastasis in head and neck cancer. *Phys Med Biol* 64:075011. <https://doi.org/10.1088/1361-6560/ab083a>
 47. Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput* 10:1895–1923. <https://doi.org/10.1162/089976698300017197>
 48. Ghasemi A, Zahediasl S (2012) Normality tests for statistical analysis: a guide for non-statisticians. *Int J Endocrinol Metab* 10:486–489. <https://doi.org/10.5812/ijem.3505>
 49. An C, Park YW, Ahn SS, Han K, Kim H, Lee SK (2021) Radiomics machine learning study with a small sample size: single random training-test set split may lead to unreliable results. *PLoS One* 16:e0256152. <https://doi.org/10.1371/journal.pone.0256152>
 50. Tian Y, Komolafe TE, Zheng J et al (2021) Assessing PD-L1 expression level via preoperative MRI in HCC based on integrating deep learning and radiomics features. *Diagnostics (Basel)* 11:1875. <https://doi.org/10.3390/diagnostics11101875>
 51. Demircioğlu A (2021) Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. *Insights Imaging* 12:172. <https://doi.org/10.1186/s13244-021-01115-1>
 52. 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>
 53. Guyon I, Hur AB, Gunn S, Dror G (2004) Result analysis of the NIPS 2003 feature selection challenge. In: *Advances in Neural Information Processing Systems* 17. MIT Press, Cambridge, pp 545–552
 54. Bommert A, Sun X, Bischl B, Rahnenfuehrer J, Lang M (2020) Benchmark for filter methods for feature selection in high-dimensional classification data. *Comput Stat Data Anal* 143:106839. <https://doi.org/10.1016/j.csda.2019.106839>
 55. Xiao B, He N, Wang Q, et al (2021) Stability of AI-enabled diagnosis of Parkinson's disease: a study targeting substantia nigra in quantitative susceptibility mapping imaging. *Front Neurosci* 15:760975. <https://doi.org/10.3389/fnins.2021.760975>
 56. Demircioğlu A (2022) The effect of preprocessing filters on predictive performance in radiomics. *Eur Radiol Exp* 6:40. <https://doi.org/10.1186/s41747-022-00294-w>
 57. Naglah A, Khalifa F, Khaleel R et al (2021) Novel MRI-based CAD system for early detection of thyroid cancer using multi-input CNN. *Sensors (Basel)* 21:3878. <https://doi.org/10.3390/s21113878>
 58. Liu Y, Sangineto E, Bi W, et al (2021) Efficient training of visual transformers with small datasets. In: *Advances in Neural Information Processing Systems* 34, pp 23818–23830
 59. Collins GS, Reitsma JB, Altman DG, Moons KGM (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg* 102:148–158. <https://doi.org/10.1002/bjs.9736>
 60. Park JE, Kim D, Kim HS et al (2020) Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. *Eur Radiol* 30:523–536. <https://doi.org/10.1007/s00330-019-06360-z>
 61. van Griethuysen JJM, Fedorov A, Parmar C et al (2017) Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 77:e104–e107. <https://doi.org/10.1158/0008-5472.CAN-17-0339>
 62. Demšar J, Curk T, Erjavec A et al (2013) Orange: data mining toolbox in Python. *J Mach Learn Res* 14:2349–2353
 63. Park JE, Park SY, Kim HJ, Kim HS (2019) Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. *Korean J Radiol* 20:1124–1137. <https://doi.org/10.3348/kjr.2018.0070>
 64. Galavis PE (2021) Reproducibility and standardization in radiomics: are we there yet? *AIP Conference Proceedings* 2348, 020003. <https://doi.org/10.1063/5.0051609>
 65. Moummad I, Jaudet C, Lechervy A et al (2021) The impact of resampling and denoising deep learning algorithms on radiomics in brain metastases MRI. *Cancers (Basel)* 14:36. <https://doi.org/10.3390/cancers14010036>
 66. Müller-Franzes G, Nebelung S, Schock J et al (2022) Reliability as a precondition for trust—segmentation reliability analysis of radiomic features improves survival prediction. *Diagnostics (Basel)* 12:247. <https://doi.org/10.3390/diagnostics12020247>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.