Progress in Biomedical
Engineering

TOPICAL REVIEW • OPEN ACCESS

Tackling the small data problem in medical image classification with artificial intelligence: a systematic review

Progress in Biomedical Engineering

OPEN ACCESS

RECEIVED
25 October 2023

REVISED
15 April 2024

ACCEPTED FOR PUBLICATION
30 May 2024

PUBLISHED
17 June 2024

TOPICAL REVIEW

Tackling the small data problem in medical image classification with artificial intelligence: a systematic review

Stefano Piffer[1,2,*], Leonardo Ubaldi[1,2], Sabina Tangaro[4,5], Alessandra Retico[3] and Cinzia Talamonti[1,2]

[1] Department of Experimental and Clinical Biomedical Sciences, University of Florence, Florence, Italy
[2] National Institute for Nuclear Physics (INFN), Florence Division, Florence, Italy
[3] INFN, Pisa Division, Pisa, Italy
[4] Department of Soil, Plant and Food Sciences, University of Bari Aldo Moro, Bari, Italy
[5] INFN, Bari Division, Bari, Italy
[*] Author to whom any correspondence should be addressed.

E-mail: stefano.piffer@unifi.it

Keywords: small data, artificial intelligence, medical imaging, transfer learning, data augmentation, classification

Supplementary material for this article is available online

Abstract

Though medical imaging has seen a growing interest in AI research, training models require a large amount of data. In this domain, there are limited sets of data available as collecting new data is either not feasible or requires burdensome resources. Researchers are facing with the problem of small datasets and have to apply tricks to fight overfitting. 147 peer-reviewed articles were retrieved from PubMed, published in English, up until 31 July 2022 and articles were assessed by two independent reviewers. We followed the Preferred Reporting Items for Systematic reviews and Meta-Analyse (PRISMA) guidelines for the paper selection and 77 studies were regarded as eligible for the scope of this review. Adherence to reporting standards was assessed by using TRIPOD statement (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis). To solve the small data issue transfer learning technique, basic data augmentation and generative adversarial network were applied in 75%, 69% and 14% of cases, respectively. More than 60% of the authors performed a binary classification given the data scarcity and the difficulty of the tasks. Concerning generalizability, only four studies explicitly stated an external validation of the developed model was carried out. Full access to all datasets and code was severely limited (unavailable in more than 80% of studies). Adherence to reporting standards was suboptimal (<50% adherence for 13 of 37 TRIPOD items). The goal of this review is to provide a comprehensive survey of recent advancements in dealing with small medical images samples size. Transparency and improve quality in publications as well as follow existing reporting standards are also supported.

1. Introduction

Data-driven intelligent models have gained immense popularity in recent years, achieving satisfactory performance. The essence behind these achievements is that the behavior in unknown domains can be accurately estimated by quantitatively learning the latent patterns behind the data from sufficient training samples [1, 2].

Researchers nowadays are capable of designing and developing network structures with even more and wider layers than before also thanks to the availability of much more powerful computational resources. The trend of artificial neural networks points towards the idea that deeper or more complicated networks perform better. However, these techniques are built up on the assumption of sufficiently large data samples for appropriate model training, i.e. Big Data. Usually, the term Big Data indicates a massive volume of data that is too large or complex to be effectively analyzed using traditional software [3, 4].

In numerous real-world applications, the number of samples in a dataset can be relatively limited, constrained by the complexity, ethnicity, high cost or difficult to obtain in practical, leading to sharply decreases the performance of deep learning models. This is the main restriction of the deep learning models: they need tens of thousands of well-labeled samples for training. This Small Data challenge would call for a completely different approach from the existing Big Data one, and the axiom 'the deeper and wider we go, the better the performance' is no longer as robust [3]. The limited quantity of available data prevents the use of large models: indeed, training smaller models is a safer choice since they are less prone to overfit data. Very large models, if not properly regularized, tend to memorize the whole dataset causing serious overfitting and a poor generalization ability of the model [5]. In fact, the small data challenge is not only about the size of the training database in absolute terms and therefore when the train data is deficient the learned feature representations are limited and the model only fits well on train data. But it is essential to contemplate the small data issue in relative terms with respect to the complexity of the model to be trained. A large, deep and complex learning algorithm with millions of free parameters to optimize can obtain an effective knowledge of the available dataset achieving good train performance, albeit at the expense of heavily parameterizing the available data and loosing model generalizability.

Another aspect that needs to be brought into view concerns the quality of the data. In the clinical context, only expert physicians can give high-quality sample annotations, and such large amounts of annotated data will inevitably be laborious, costly and time-consuming. This prevents the creation of sufficiently large samples in most cases [4, 6]. In this perspective, small sample size issue is of particular interest when neural networks are applied to medical images, including MRI, CT, dose distributions, ultrasounds, and histopathological images, which often have limited sample size restricted by the availability of the patient's population, scarcity of annotated datasets and experts' labeling. In general, for medical images, high-quality annotated datasets are scarce and require specialized medical knowledge, standardized protocols and considerable time and effort. For this purview, labeling of data by domain experts is still one of the key issues and often it may take more time and effort than the algorithm development itself. Moreover, the intrinsic heterogeneity of retrospective data accumulated in daily clinical practice creates a trade-off between the quality and the dataset sizes, ranging from a few dozens to a few hundreds of patients [7, 8].

Moreover, constructing sufficiently large data sets in the field of medical imaging is difficult due to the patient privacy and regulations. For this reason, starting multicenter studies is often a difficult path to take and individual clinical centers try to train, validate and test artificial intelligence algorithms with the few available data. But a small sample size from a single study database produces fundamental limits. Deep learning techniques generally require more than a million samples to train without overfitting. However, another important aspect present in clinical studies must also be emphasized. In this context, rare diseases are often studied and therefore lack data per se, or they have to deal with classes or categories that are numerically very unbalanced [9]. Consequently, many deep learning researchers agree that a small sample size is insufficient to test the effectiveness of the proposed method. In recent years, some international competitions have released rich labeled medical images, which provides a potential data source to train models specific to medical applications.

The small data issue can be mainly solved with two approaches: data augmentation-based and transfer learning/domain adaptation-based, respectively. These methods try to expand the data volume but in a different fashion. The first method is based on the generation of new synthetic data from the available data while the second one resorts on knowledge learned from other domains. These methods could effectively improve the results and reduce the data size requirement in order to overcome the Small Data challenge. They are illustrated in detail below.

Data augmentation. The data augmentation-based strategy aims to synthetically and artificially increase the number of available samples for training deep learning models miming the distribution of the original dataset, providing more general information from the dataset to solve the small data problem. It is a data preprocessing method and a type of regularization which can effectively improve the performance of model reducing the possibility of overfitting [10, 11].

Two very simple augmentation processes are generally employed: gray level disturbance and shape disturbance. In the first case, Gaussian noise or something similar is added to the original images. In the second one, the data is increased by oversampling images with translations, rotations, brightness modification, rescaling, flipping, shearing or stretching and other affine transformations. In general, the idea behind these operations is that they will assist the learning algorithm to acquire more comprehensive and robust features which will then be useful in conditions where the data could be incomplete and/or noisy, favoring generalization.

One such more objective and promising technology that recently has been introduced for data augmentation, are the generative adversarial network (GAN) which involves generative models and adversarial learning [12, 13]. The GAN attempts to approximate the true data distribution through a minimax game between two subnetworks in competition with each other, called the discriminator and the generator. The generator attempts to create data samples as similar as possible to the true data while the discriminator seeks to distinguish true from fake-generated samples. The two subnetworks evolve together during training; the generator tries to deceive the discriminator by improving its output more and more, in other words, it learns to approximate better and better the distribution of the original data. Thereby new completely synthetic data samples can be generated and used for training in the main task. In general, as a generative model, a well-trained GAN is used to provide additional fake and synthetic samples that has the same distribution with the original training data [14–17].

Transfer learning. Another possible way to solve the small sample size problem is transfer learning, that is to use a pre-trained network, which cleverly applies knowledge gained from a source domain to facilitate the learning problem in a partially related or unrelated target domain. Transfer learning provides an effective framework for deep learning with small datasets; it pretrains a model by using existing massive dataset and then uses the trained model either as an initialization or as it is for a new task [18–20].

The idea is to initialize the neural network with the weights trained from some previous task and fine-tune parameters within the current task when the current task has insufficient training data. This approach provides a reasonable initial state and may speed up training, slightly different form the traditional learning process where it tries to learn each task from scratch. There may be three different approaches to reuse the parameters (weights and biases) of pre-trained network: (1) reusing the parameters in pre-trained deep neural network directly to initialize the new network and fixing without retraining, called freezing. (2) Reusing the parameters in pre-trained deep neural network directly to initialize the new network and fine-tuning the parameters using target domain data, called fine-tuning. (3) Initializing network parameters randomly and tuning parameters using target domain data, called random initialization and training [1].

The source domain can pertain to a connected sphere of the target task as well as to a completely different one. As a matter of fact, most studies have made use of models pretrained from the large-scale ImageNet database [21], containing 1.2 million natural images. These models trained from the ImageNet have a strong capability for feature extraction. Thus, they are suitable to be transferred to other context having small number of image data and can produce significant advanced performances better than shallow algorithms. Such a strategy reduces the need and effort to recollect a large training data, saving data resources and training time. Transfer learning cloud be very effective in the field of medical images where pretraining can mitigate the drawback of having a very large labeled datasets and can prove very useful in building complex and robust models. In general, the use of deep neural networks even with small data samples can occur thanks to the pre-training on data-rich domains that share affinities in statistical properties with the target dataset [22–24].

The aim of this work, and the related research question, is to present a systematic review to provide an overview of the state of the art of deep learning research for clinical applications on small samples and to highlight the different strategies for working in this scenario. Specifically, we sought to describe the study characteristics, and evaluate the methods and quality of reporting and transparency of deep learning studies that compare diagnostic algorithm performance with the ground truth.

2. Methods

2.1. Preferred Reporting Items for Systematic reviews and Meta-Analyse Prisma

This manuscript has been prepared according to the guidelines and a checklist is available in the supplementary material [25].

2.2. Literature search and inclusion criteria

We performed a comprehensive search by using free text terms for various forms of the keywords ‘small’, ‘data base’ and ‘deep learning’ to identify eligible studies. PubMed MEDLINE database was thoroughly searched to identify original research articles that investigated the performance of AI algorithms analyzing small medical images samples. We used the following search query: (‘small’ OR ‘limited’) AND (‘sample’ OR ‘samples’ OR ‘database’ OR ‘databases’ OR ‘dataset’ OR ‘datasets’ OR ‘data sample’ OR ‘data samples’) AND (‘medical images’ OR ‘medical imaging’) AND (‘artificial intelligence’ OR ‘radiomics’ OR ‘machine learning’ OR ‘deep learning’) AND (‘classification’ OR ‘prediction’ OR ‘clustering’). PubMed search engine was questioned without imposing time filters (literature search update until 31 July 2022).

We selected publications for review if they satisfied several inclusion criteria: a peer reviewed scientific report of original research; English language; assessed a deep learning algorithm applied to a clinical problem in medical imaging; application of the AI techniques on declared small datasets; and compared algorithm performance with the ground truth.

We included studies when the aim was to use medical imaging for predicting absolute risk of existing disease or classification into diagnostic groups (e.g. disease or non-disease). In machine learning, regression and classification are closely related concepts in that they both involve making predictions from data and they both play crucial roles in medical image analysis. Even though they can be used together in a cascaded or integrated approach, these two procedures differ in terms of their objectives and the nature of the output they produce. Regression aims to predict a continuous numerical value as the output. In the context of medical image analysis, they provide quantitative information about various aspects of patient health like tumor size, bone mineral density, blood flow quantification, etc. Classification, on the other hand, focuses on assigning inputs to predefined categories or classes. In medical image analysis, these classes might represent different diseases or conditions (normal or abnormal, malignant or benign). Fundamentally, regression is about predicting a quantity and classification is about predicting a label. Since in the final analysis they can be considered as two very distinct tasks and that the choice between them depends on the nature of the assignment and the information required from the analysis, we have decided to focus only on the classification task to narrow the research and be able to obtain a more homogeneous set of results than allows us to reach the most rigorous assessments.

We defined medical images as radiologic images and other medical photographs (e.g. endoscopic images, retinal images, pathologic photos, and skin photos) and did not consider any line art graphs that typically plot unidimensional data across time, for example, electrocardiogram and A-mode ultrasound. Case reports, review articles, editorials, letters and comments were left out. Exclusion criteria included also AI algorithms that performed image-related tasks other than direct diagnostic decision-making, such as image segmentation, databases description and manage data preprocessing.

2.3. Screening of collected studies

After removal of clearly irrelevant records, two reviewers independently screened abstracts for potentially eligible studies. Abstracts with any degree of ambiguity or that generated differences in opinion between the two reviewers were re-evaluated at a consensus meeting, for which a third reviewer was invited.

The admissibility of the full text articles was then assessed by the same reviewers as before who will then extract the data from the study reports. After this second screening, articles belonging to one of the following categories were excluded: methodological works, object detection tasks, focus on explainability and out of the topic.

2.4. Adherence to reporting standards—transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)

We evaluated the quality of the studies according to the TRIPOD statement [26]. This statement rates the transparency of the reporting of a prediction model study regardless of the survey methods used and in all medical settings [27]. It is composed of a 22 items checklist (37 total points when all subitems are included), which analyzes the development, validation, or the updating of a prediction model, whether for diagnostic or prognostic purposes. The aim was to assess whether studies broadly conformed to reporting recommendations included in TRIPOD, and not the detailed granularity required for a full assessment of adherence [28].

2.5. Data synthesis and analysis

Aware of heterogeneity of specialties, metrics and outcomes, we reported in table 1 the basic qualitative and quantitative characteristics such as anatomical region, AI technique, sample size, number of classes, best performance, type of images, programming language and sharing code and database.

Two-sided Mann–Whitney–Wilcoxon statistical test was conducted with Bonferroni correction and an alpha value of 0.05 was used to determine significance.

3. Results

3.1. Study selection

Our electronic search carried out considering only the filter 'titles and abstracts', which was last updated on 31 July 2022, retrieved 147 records. Of the 147 initially collected studies, we assessed 105 full text articles; 28 were excluded, which left 77 works for analysis (figure 1).

Compared to existing reviews, our work is original and contemporary, and faces a very important hot topic of artificial intelligence in the field of medical images. In fact, none of the reviews that emerged from the search query identified by us detailed the problem of the small dataset. In particular, most of the reviews

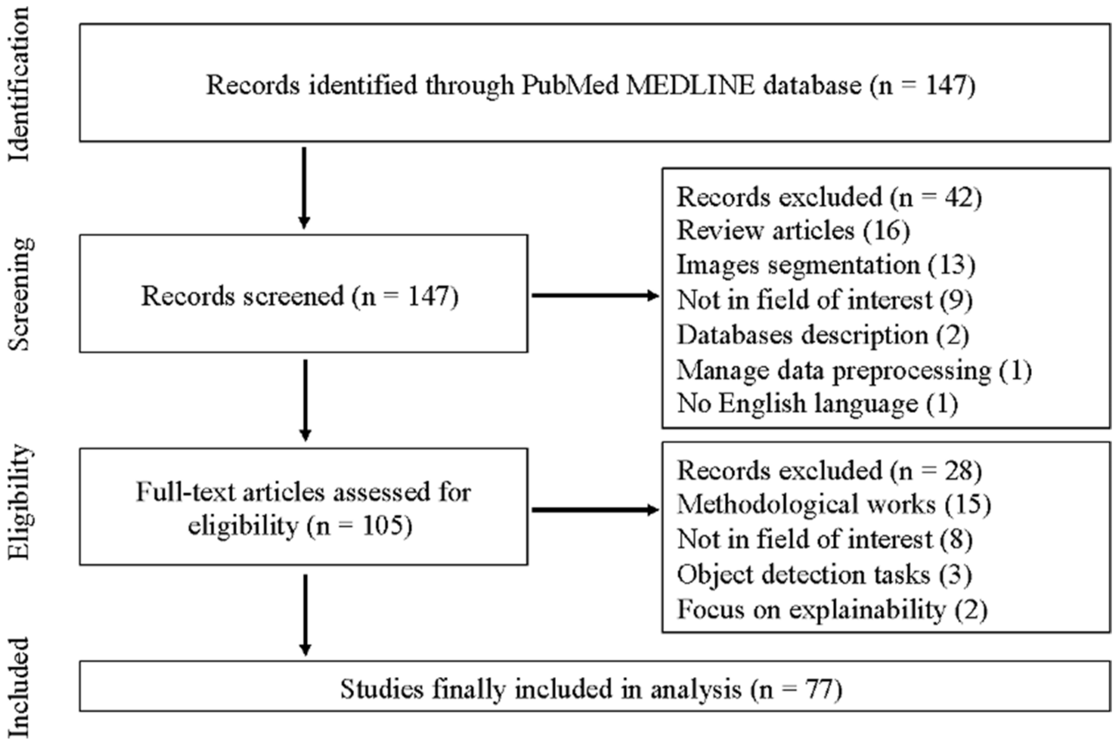
Identification
Records identified through PubMed MEDLINE database (n = 147)
Screening
Records screened (n = 147)
Records excluded (n = 42)
Review articles (16)
Images segmentation (13)
Not in field of interest (9)
Databases description (2)
Manage data preprocessing (1)
No English language (1)
Eligibility
Full-text articles assessed for eligibility (n = 105)
Records excluded (n = 28)
Methodological works (15)
Not in field of interest (8)
Object detection tasks (3)
Focus on explainability (2)
Included
Studies finally included in analysis (n = 77)

Figure 1. Flow-chart. Flow-chart of article selection based on PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines.

did not handle the problem of small dataset, and only six named this issue, without however entering in detail and discussing the topic thoroughly. In these cases, the authors of the reviews faced how artificial intelligence is applied to a specific task, simply summarizing in the conclusions that one of the main drawbacks of the analyzed studies concern the use of small databases for the training of the AI algorithms.

3.2. General characteristics

Table 1 summarizes the basic characteristics of the 77 studies. All of them are about the development and the validation of a prediction model, specifically, 75 (97%) publications deal with diagnostic model and only 2 (3%) with prognostic model. Most of the works make use of deep learning techniques (86%), only 6% applies purely machine learning techniques and 8% mix both methodologies.

The top five imaging modalities are x-ray 23/77 (30%), MRI 19/77 (25%), CT 18/77 (23%), histological 9/77 (12%), and ocular images 5/77 (6%). The remaining types concern ultrasound, endoscopic, PET and SPECT images (figure 2(a)). Zooming on the first three categories, x-ray images take care of lungs (12), breast (8), skeleton (2) and adenoid (1); MR images focus on brain (13), prostate (3), knee (2) and liver (1); CT images pay attention on lungs (10), H&N (2), colon (2), liver (2), heart (1) and brain angiography (1). As regards the number of samples in the databases, they present a distribution with an average population of 16 600 ± 45 700 samples (mean ± one standard deviation), a minimum of 16 and a maximum of 299 000 (figure 2(b)). Most of the studies develop AI techniques by exploiting the clinical images of the anatomical regions most investigated in the clinic and therefore with the greatest probability of finding adequate databases: brain, breast, lung (figure 2(c)). Furthermore, as can be expected, given the scarcity of data in small samples and the difficulty of the tasks, more than 60% of the authors perform a binary classification (figure 2(d)). Concerning reproducibility, data are public and available in 47 studies. In 25 analysis the collected data are private and 7 operate over both types of databases. 50% of the studies managed only one repository, 31% acted on 2, 10% employed 3 databases and the rest of the publications more than three. Additional plots relative to the quantity of available data with respect to the anatomical region, the imaging technique and the dataset origin can be found in the supplementary materials (supp. figure 1).

To solve the small data issue transfer learning techniques, basic data augmentation and GANs are applied in 75%, 69% and 14% of cases, respectively. All three methodologies are exploited simultaneously in only 8 studies, while 26 used none of these techniques. The two main metrics used are accuracy and area under curve (AUC). The first was used in 65/77 studies to evaluate the performance of the algorithm on the test set, obtaining an average value of 0.90 ± 0.11, while the second was used in 48/77 works with an average value of 0.90 ± 0.10.

Table 1. Characteristics of included studies. TL (transfer learning), DA (data augmentation), CV (cross-validation), T:V:T (training:validation:test).

First Author	Publication year	TL	Classic DA	Advanced DA	Available data	Augmented available data	CV	T:V:T split (%)^{a}	Test Accuracy	Test AUC	# of classes	Imaging modality	Anatomical region	Database type	#of used database	External Test	Sharing code
Swati *et al* [2]	2019	X			233		5-fold		0,948		3	MRI	Brain	Public	1		X
Abbasi *et al* [29]	2021	X	X		26 431	31 231	5-fold		0,823	0,889	2	Retinal images	Eyes	Public	2		
Ali Khan *et al* [30]	2020		X		253			73:19:08	1,000	1,000	2	MRI	Brain	Public	1		
Romero *et al* [1]	2020	X	X		112 120			70:10:20		0,950	2	XRay	Lung	Public	4		
Horry *et al* [20]	2020	X	X		2368	28 560		80:20:00			2	Ultrasound, XRay, CT	Breast	Public	4		
Alzubaidi *et al* [31]	2021	X	X		143 243	343 243		80:20:00	0,975		4	Histologic	Skin	Public	2		X
Wodzinski *et al* [32]	2020	X	X		174		5-fold		0,740		2	MRI	Brain	Private	1		
Hertel *et al* [33]	2021	X	X		31 595	37 914		90:10:00	0,940	0,982	3	XRay	Lung	Public	5		
Baydilli *et al* [34]	2020				263		10-fold	75:10:15	0,969	0,925	5	Histologic	White Blood Cells	Public	1		
Li *et al* [35]	2019	X	X		15 573	140 157	10-fold	85:05:10	0,973	0,995	4	Retinal images	Eyes	Mix	3		
Xia *et al* [36]	2018	X			299 096			90:10:00	0,843	0,919	2	Histologic	Lymph node metastases	Public	2		
Shen *et al* [37]	2020	X	X		668			73:10:17	0,956		3	XRay	Adenoid	Private	1		
Feng *et al* [38]	2018				58 000		10-fold		0,983		2	Histologic	Breast	Public	1		
Liu *et al* [39]	2020	X		X	24 000	27 000		75:13:12	0,882	0,931	2	Histologic	Brain	Mix	2		X
Levine *et al* [14]	2020	X	X	X	1022		10-fold		0,920	0,992	10	Histologic	Different tumors	Public	2		X
Ahn *et al* [40]	2020	X	X		13 986		5-fold	65:35:00	0,830	0,831	2	Histologic	Skin	Public	3		
Gheshlaghi *et al* [17]	2021	X	X	X	13 338	14 838		70:30:00	0,900		2	Histologic	Breast	Public	1		
Montoya *et al* [41]	2018				105			33:08:59	0,991		3	CT	Brain	Private	1		
Xia *et al* [42]	2020				373			65:35:00		0,900	2	CT	Lung	Private	2		X
Liang *et al* [43]	2021	X	X		100	640		80:20:00	0,910	0,910	2	MRI	Brain	Public	1		
Huynh *et al* [44]	2016	X			607		5-fold			0,860	2	XRay	Breast	Private	1		
Zhang *et al* [45]	2021				1870				0,570	0,597	6	MRI	Brain	Public	2		

(Continued.)

Table 1. (Continued.)

Hu *et al* [46]	2020	X			499		5-fold	55:45:00	0,710	0,753	2	MRI	Brain	Public	2		
Dai *et al* [47]	2021	X	X		1714			80:10:10	0,949	0,981	3	MRI	Knee	Private	1		
Apostolopoulos *et al* [48]	2020	X	X		216	2216	10-fold		0,745	0,810	2	SPECT	Heart	Private	1		
Bien *et al* [49]	2018	X	X		2287				0,913	0,993	3	MRI	Knee	Mix	2	X	X
Fu *et al* [50]	2020	X	X		1091			60:20:20	0,951	0,978	3	Histologic	Skeletal	Public	1		
Chougrad *et al* [51]	2018	X	X		6116		5-fold		0,982	0,990	2	XRay	Breast	Public	3		
Hu *et al* [52]	2021	X	X		217		5-fold		0,920	0,990	2	MRI	Prostate	Private	1		
Shi *et al* [15]	2020	X	X	X	1937		10-fold	84:09:07	0,915	0,953	2	US	Thyroid	Private	1		
Ye *et al* [53]	2020	X	X		650		10-fold		0,920	0,959	2	CT	H&N	Private	1		
Zhou *et al* [54]	2021	X	X		616			75:25:00	0,825		2	CT	Liver	Private	1		
Yi *et al* [55]	2019	X	X		3034			70:10:20		1,000	2	XRay	Breast	Public	1		
Mutasa *et al* [56]	2018		X		10 289		8-fold	86:10:04				XRay	Skeletal	Mix	2		
Mzoughi *et al* [57]	2020		X		284			75:25:00	0,965		2	MRI	Brain	Public	1		
Wang *et al* [58]	2017	X	X		233		10-fold				5	CT	Lung	Public	1		
Samala *et al* [59]	2020	X	X		3411	27 288	4-fold			0,830	2	XRay	Breast	Mix	2		
Yi *et al* [60]	2019	X	X		250	5760		64:12:24		1,000	5	XRay	Skeletal	Mix	2	X	
An *et al* [61]	2021	X	X		954			80:20:00		0,910	5	Retinal images	Eyes	Private	1		
Owais *et al* [62]	2020	X	X		52 471		2-fold	50:50:00	0,962		37	Endoscopic	Gastrointestinal	Public	2		X
Samala *et al* [63]	2021	X			4577		4-fold	70:10:20		0,720	2	XRay	Breast	Mix	2		
Cogan *et al* [64]	2019	X	X		8000	27 200		85:15:00	0,985	0,940	8	Endoscopic	Gastrointestinal	Public	1		
Apostolopoulos *et al* [65]	2020	X	X		3905		10-fold		0,992		2	XRay	Lung	Public	3		
Choi *et al* [66]	2017	X	X		279	10 000	5-fold	70:30:00	0,874		2	Retinal images	Eyes	Public	1	X	
Zong *et al* [67]	2020	X	X		528		10-fold	60:40:00	0,850	0,840	2	MRI	Prostate	Public	2		
Zebin *et al* [16]	2021	X	X	X	802	902	5-fold	80:20:00	0,968		3	XRay	Lung	Public	2		X
Aderghal *et al* [22]	2020	X	X		1551	184 320		70:20:10	0,920	0,940	2	MRI	Brain	Public	1		

(Continued.)

Table 1. (Continued.)

First Author	Publication year	TL	Classic DA	Advanced DA	Available data	Augmented available data	CV	T:V:T split (%)[a]	Test Accuracy	Test AUC	# of classes	Imaging modality	Anatomical region	Database type	#of used database	External Test	Sharing code
Uemura *et al* [68]	2021	X		X	333	12 407		80:20:00		0,892	2	CT	Colon	Private	1		
Oakden-Rayner *et al* [69]	2017				15 957		6-fold		0,687	0,677	2	CT	Lung	Private	1		
Nabizadeh-Shahre-Babak *et al* [70]	2021				22 232				0,998		2	XRay	Lung	Public	3		
Wang *et al* [71]	2020	X	X	X	206	48 900	3-fold		0,605	0,830	3	CT	Lung	Private	2		X
Haga *et al* [72]	2018				40		30-fold	70:30:00	0,656	0,728	2	CT	Lung	Private	1		
Fantini *et al* [73]	2021	X	X		10 880	76 800	3-fold	70:30:00	0,946		2	MRI	Brain	Private	4		
Trivizakis *et al* [8]	2019		X		130	796		57:25:18	0,830	0,800	2	MRI	Liver	Private	1		
Bahgat *et al* [19]	2021	X	X		12 933			85:15:00	0,985	0,998	4	XRay	Lung	Public	8		X
Zhang *et al* [74]	2019		X		130		2-fold		0,830	0,860	2	CT	Colon	Private	2		
Toda *et al* [75]	2021	X	X	X	66		3-fold		0,615		3	CT	Lung	Private	1		
Gatidis *et al* [11]	2015				16		10-fold	94:06:00	0,890		2	MRI, PET	Prostate	Private	1		
Sha *et al* [24]	2019				100			74:26:00	0,885	0,728	2	PET	Lung	Private	1		
Usman *et al* [76]	2022	X			197 087			80:10:10	0,890	0,870	14	XRay	Lung	Public	2		
Kaur *et al* [77]	2022				2482			80:20:00	0,994	0,999	2	CT	Lung	Public	1		
Alruwaili *et al* [18]	2022	X	X		99			70:10:20	0,895		2	XRay	Breast	Public	1		
Adedigba *et al* [10]	2022	X	X		410	18 200		80:15:05	0,998		6	XRay	Breast	Public	1		
Hashemzehi *et al* [78]	2021				6328			80:20:00	0,927		3	MRI	Brain	Public	2		
Rocca *et al* [79]	2021				30				0,933		2	CT	Liver	Private	1		
Suganyadevi *et al* [80]	2022	X			7000		5-fold	70:10:20	0,988		2	XRay	Lung	Private	2		
Ahmad *et al* [81]	2022	X	X	X	3064			60:20:20	0,963		3	MRI	Brain	Public	1		
Le *et al* [82]	2022	X	X		669		5-fold	80:20:00	0,787	0,820	3	CT	H&N	Public	2	X	
Ayana *et al* [7]	2022	X	X		21 000	28 920	nested 5-fold	70:15:15	0,990	0,999	2	Histologic, US	Breast	Public	3		
Cahan *et al* [83]	2022	X	X		358			68:12:20	0,847	0,880	2	CT	Heart	Private	1		
Ho *et al* [84]	2022	X	X		3783		10-fold	80:20:00	0,941		3	XRay	Lung	Public	5		
Muhammad *et al* [85]	2022	X	X		19 196	29 576			0,995		2	XRay, CT	Lung	Public	3		
Sanchez *et al* [23]	2022	X		X	2973		10-fold	94:06:00	0,978	0,960	2	XRay	Lung	Mix	2		
Sahoo *et al* [86]	2022	X	X		85 672			80:20:00	0,991	1,000	3	XRay, CT	Lung	Public	2		X
Zhang *et al* [87]	2022	X	X	X	7254		5-fold		1,000	1,000	2	Retinal images	Eyes	Public	3		X
Ben Ahmed *et al* [88]	2022	X	X		209	10 000		80:20:00	0,740	0,700	2	MRI	Brain	Public	1		X
Ettehadi *et al* [89]	2022				6720			60:20:20	0,975		4	MRI	Brain	Public	2		

[a] Respect to the original available data.

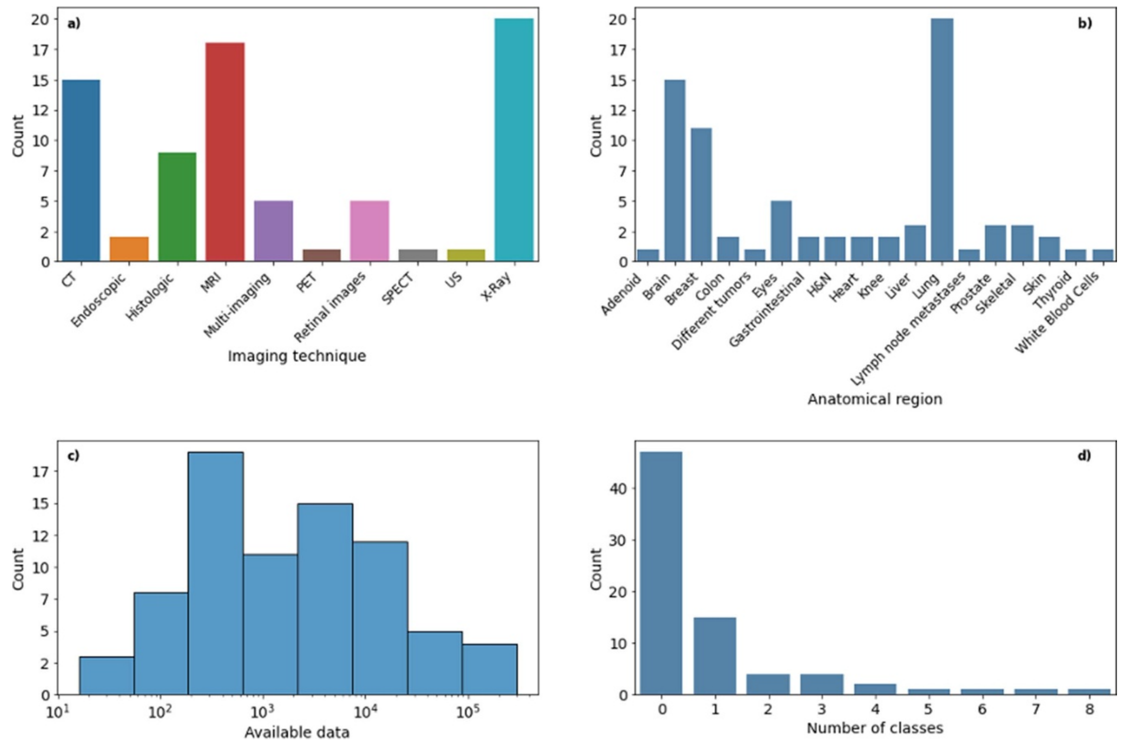

Figure 2. General characteristics. General characteristics concerning the imaging modalities (a), distribution of the number of samples in the databases (b), the most popular anatomical regions (c) and the preferred type of classification (d).

Fifty-three of 77 (69%) studies claimed in the discussion that the prediction model could have a potential clinical use (e.g. to identify high risk groups to help clinicians in decision making, or to triage patients for referral to subsequent care). Moreover, 90% of the authors declared that improvements and future research are necessary (e.g. a description of what the next stage of investigation of the prediction model should be). Relative to transparency and sharing, code (for preprocessing of data, modeling and reproducing the evaluation) is available in only 13 studies (17%). Funding was predominantly academic (45/77, 58%) and mixed with commercial supporters in 3 cases (4%). Ten studies stated they had no funding and another 19 did not report on funding.

In the following analysis, in order to better interpret the results and since most of the works take into consideration a binary classification as mentioned before, we focused only on these studies and we wanted to verify a possible increase in the performance of AI algorithms in terms of accuracy and AUC as function of publication year (figure 3). None of the data is statistically significant but a growing trend can be visually appreciated. This could be due to the growing use of transfer learning and data augmentation (figure 4). By comparing the performance metrics with respect to the use or not of these techniques, differences can be noted (figure 5). For both accuracy and AUC, if transfer learning, data augmentation or both AI techniques are exploited, the dispersion of data is more limited, both in terms of interquartile range and whisker extension. Furthermore, even if for accuracy the median values of the distributions with and without the use of the different techniques are comparable, for the AUC the difference between these values is considerable. In point of fact, the use or not of data augmentation is statistically significant ($p = 0.03$).

This analysis presents potential biases and confounders, such as different methods, different tasks or different numbers of initial data that could influence the performances, due to the presence of some limitations in the available data. Here are the assumptions for which we considered the data as consistent and coherent for comparison. It is very difficult to find a homogeneous set with a specific task but all the works examined for these plots have binary classification as a common task. Furthermore, as regards models and databases, based on what was declared by the authors, the initial databases can be considered small compared to the parameters of the models to be optimized during training, regardless of the type and imaging modalities examined.

3.3. Adherence to reporting standards

Adherence to reporting standards less than 50% is present in 13 of 37 TRIPOD items (figure 6). Overall, publications adhered to between 52% and 88% of the TRIPOD items: median 68%, interquartile range

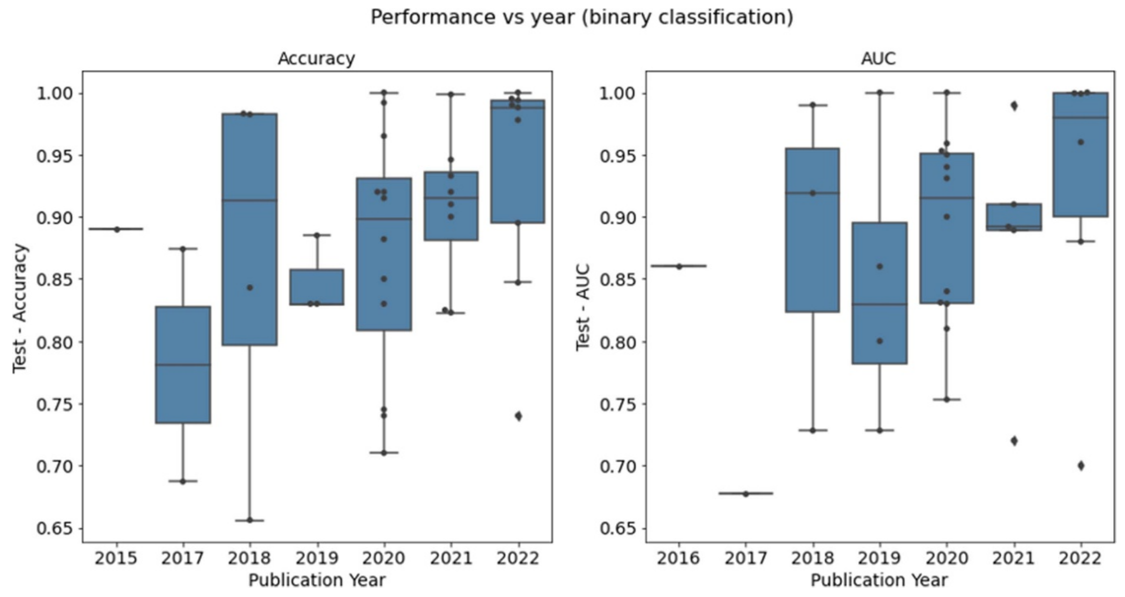

Figure 3. Performance of AI algorithms. Performance of AI algorithms in terms of accuracy (left) and AUC (right) as function of publication year for binary classification studies.

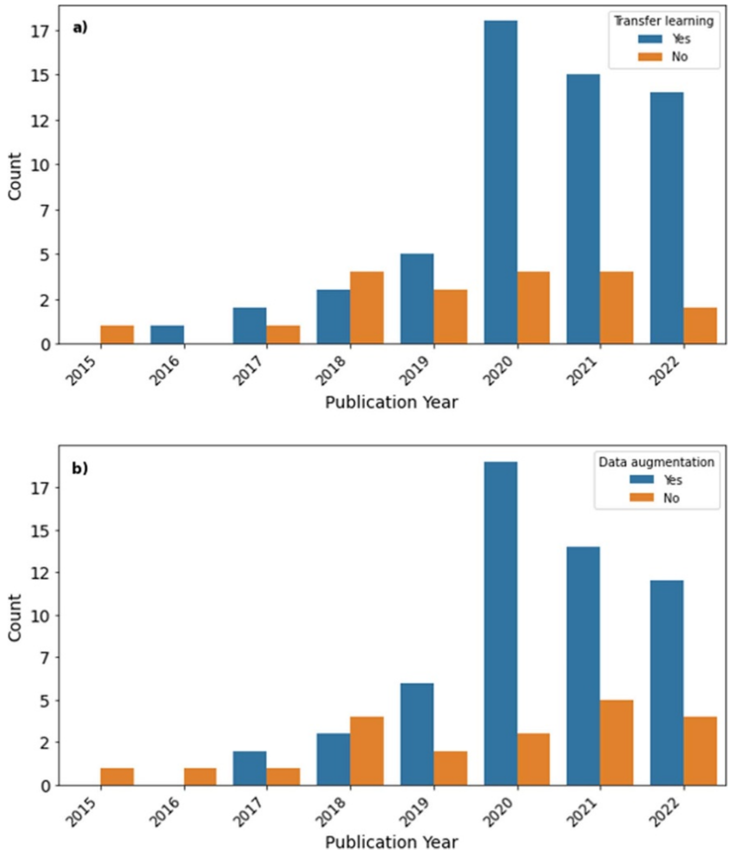

Figure 4. Use or not of AI techniques. Use of transfer learning (a) and data augmentation (b) as function of publication year.

61%–71%, confidence level at 5 and 95% are 55 and 81%, respectively, corresponding to two studies below the 5% threshold and three studies above the 95% threshold.

Two items deserve deep comment: number 1 (identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted) with an adherence of 3% and number 16 (report performance measures with confidence intervals for the prediction model) with an adherence of 29%. In the first case such low adherence has found because in the title the authors have

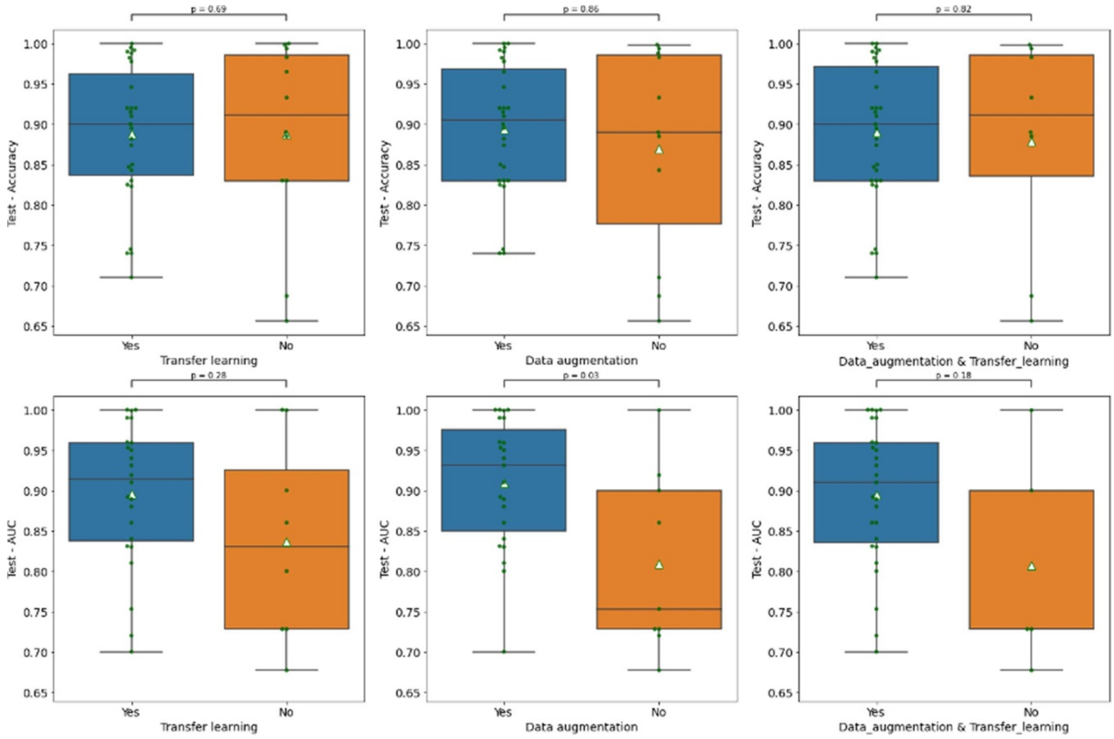

Figure 5. AI performances. AI performances (accuracy top, AUC bottom) for binary classification studies; with and without transfer learning (first column), with and without data augmentation (second column), with and without both techniques (third column).

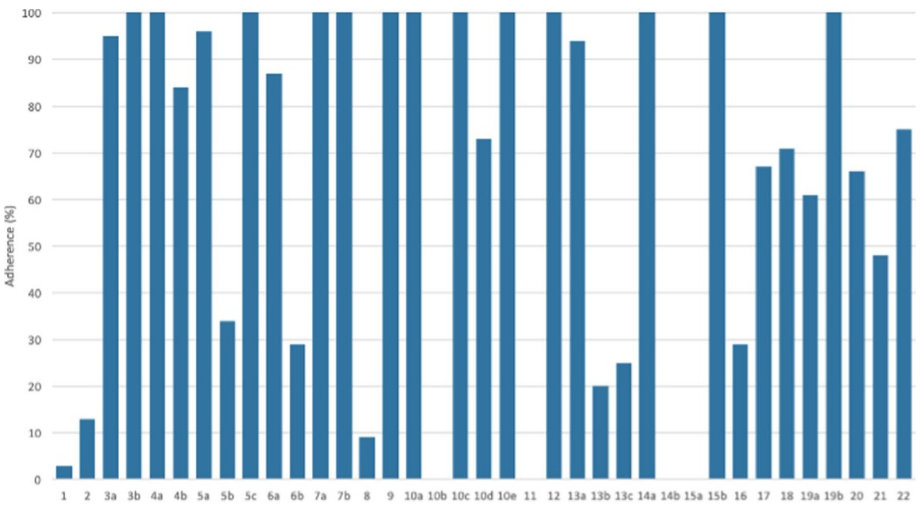

Figure 6. Adherence to reporting standards. Completeness of reporting of individual TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) items.

not reported the words development, validation, incremental/added value (or synonyms). While in the second one, the confidence interval (or standard error) of the discrimination measure and/or the measures for model calibration are often not indicated.

The full results of TRIPOD adherence assessment form for this study are available in the online supplement materials.

For the moment, quantity and quality have not helped to improve performances (figure 7). On one hand, perhaps the quality of the data needs to be boosted and/or even if a large database is available, it is not guaranteed to obtain excellent performance because it probably contains greater heterogeneity by representing the real variability in a more objective way. On the other, having a high TRIPOD index is not a guarantee of having good performances since it mainly evaluates the reliability and transparency of the studies. Additional plots relative to the performances with respect to the quantity (available data) and the

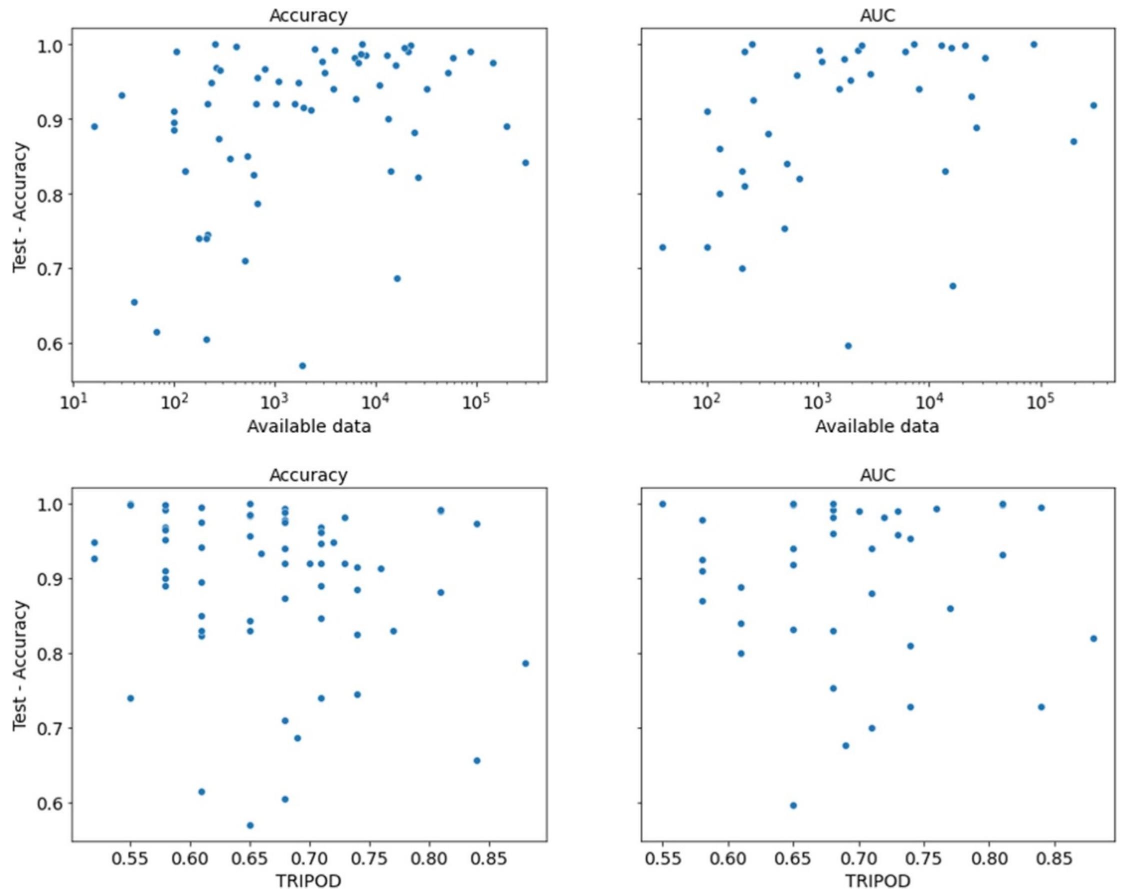

Figure 7. Quantity & quality. AI performances with respect to the quantity (Available data) and quality (TRIPOD) of the data.

quality (TRIPOD index) by anatomical region, imaging technique and dataset origin can be found in the supplementary materials (supp. figures 2 and 3).

4. Discussion

We have conducted an appraisal of the methods and adherence to reporting standards. These studies are constantly increasing and are pushing more and more to introduce AI algorithms into clinical practice as quickly as possible. The potential consequences for patients for immature implementation of these systems without a rigorous evidence base could be catastrophic. For the moment, the efforts should focus on improving design, validation, transparency and sharing [82].

All the selected works declare that the database at their disposal was small and therefore limited for an optimal achievement of their objective. But as can be seen from table 1, certain databases are difficult to classify as small in absolute terms having more than 100 000 data. It is therefore essential to declare the term ‘small’ in relative terms with respect to the number of free parameters to be optimized. In this way it is more evident how difficult it is the task of training a complex model prone to overfit the data and without an appropriate regularization [90].

Working with small databases there is the risk of creating a bias in the optimized model due precisely to the few samples available and this negatively affects its generalizability and reliability. Even if the algorithm is tested on a subset of data not used during training, if not handled properly, when testing the algorithm on an external dataset this can lead to a poor performance [5, 91, 92].

The works we encountered are retrospective studies and only four explicitly stated that they have carried out an external validation of the developed model, meaning using a completely independent database compared to the previous one, with another patients’ distribution, coming from a different geographical region or using a real hospital database. For this reason, they should be considered only a proof of concept and there is still a long way to go before being able to arrive at an effective clinical implementation. There are comparisons of the AI performance with respect to clinicians, but unfortunately they are still minimal and the very good performances obtained in silico may not lead to an effective clinical benefit, such as an unacceptably high false positive rate. Entering in more detail in this area, one should verify or at least be aware of how clinical ground truths are defined. First, because there is variability between intra and inter expert clinicians and the most likely value would be that generated by a suitably large sample of experts to ensure reliability. Second, because the inclusion of non-experts is starting to take hold, especially in segmentation tasks. Such a tendency can lower the average human performance and potentially make the AI algorithm perform better than it otherwise might [93]. In this perspective, particular attention should be paid if public databases are used; however useful and sometimes essential, before throwing yourself headlong into training AI algorithms, it is better to inquire in detail about how the database was built and how the ground truths were obtained. In addition to the quantity, the quality and certifiability of the data should also begin to be considered a must.

Developing AI systems employing tens of thousands of training samples leads to onerous investments since high level knowledge is required to prepare such data. Therefore, designing AI algorithms under small amounts of quality data with high accuracy is of great significance and an important direction of current artificial intelligence research. To overcome the main drawbacks and pitfalls in this field, reliable and efficient strategies must be considered and applied [31, 47, 59].

With medical images, the dimensional differences between 2D and 3D medical data present several challenges and aspects, especially when training neural networks for medical image analysis. The most obvious consideration concerns the fact that a single 3D volume can be seen as a stack of several hundreds of 2D images, which can lead to a significant increase in the amount of available data. In addition, other aspects must be taken into consideration which concern the intrinsic distinctions in the quantity of information, spatial relations and complexity.

The choice of patient classification based on 2D images, as opposed to 3D volumes, is a strategy that is taking root and spreading in literature [2, 94, 95]. In many medical settings, the acquired and readily available images are typically in the form of 2D slices with a notable slice gap. This practice is prevalent in various imaging modalities such as CT and MRI. So long as 3D volumes encompass a stack of consecutive slices, the main strategic advantage of adopting a 2D-based approach is the ability to leverage a larger pool of training samples for deep neural networks. Instead of considering the entire 3D volume for each patient, researchers can extract transversal 2D slices. This extraction process enables the generation of multiple training samples from a single patient, equal to the number of transversal slices available for analysis. When counting the overall number of training samples, it is feasible to go from several tens in the original dataset to thousands after slicing the patients. Consequently, the dataset for training the ML models is significantly enriched, enhancing the ability of the model to generalize and to learn from diverse perspectives within each patient's imaging data. This approach addresses the potential limitations associated with limited datasets, especially in the context of medical imaging where obtaining labeled data for training can be challenging. The increased number of training samples contributes to the robustness and adaptability of complex ML models, such as deep neural networks, contributing to more accurate and clinically relevant outcomes.

Certainly, the above-mentioned approach is cunning, but other aspects must be kept in mind if someone chooses to disarticulate a 3D volume into 2D. The sheer volume of data in 3D is significantly larger than its 2D counterpart and this can pose challenges in terms of storage, computational resources and complexity, and training time. But there are also intrinsic distinctions in the amount of information and spatial relations associated with 2D and 3D modalities. Volumetric medical images preserve spatial relationships and context that may be lost in 2D representations and therefore may pose challenges in capturing the continuous and detailed information necessary for certain medical tasks. Neural networks trained on 3D data can potentially catch more comprehensive information about the three-dimensional structure of anatomical features, leading to better performance in tasks requiring spatial understanding. Addressing the dimensional differences between 2D and 3D medical data requires careful consideration of the specific task, available resources, and the nature of the medical imaging data. Developing effective ML and DL architectures and data augmentation strategies is crucial for achieving optimal performance in medical image analysis tasks.

As the systematic review revealed, researchers rely mostly on data augmentation and transfer learning. Inherently to the first solution to enrich the dataset via the augmentation strategies, it should be underlined how the use of affine transformations to create new (similar) versions of existing samples without adding any morphological variations cannot fully resolve the overfitting problem. The generated images become much correlated to each other offering modest improvement for further generalization over unseen samples. On the other hand, the spread of GANs with their astounding abilities can help to address overfit, creating morphological variations in augmented samples while preserving the key characteristic. Ahmad *et al* [81] proposed a framework based on unsupervised deep generative neural networks to solve the need for a large amount of medical images. They combined two generative models in the proposed framework: variational autoencoders and GANs. Artificially generated brain tumor images were used to augment the real and available images during the classifier training performed with ResNet50. By using brain tumor images generated artificially, classification average accuracy improved from 72.63%, without classic augmentation and generative images, to 96.25%, with classic augmentation and generative images. Wang *et al* [71] proposed an automatic classification system for subcentimeter pulmonary adenocarcinoma, combining a homemade convolutional neural network (CNN) and a GAN. For GAN-based image synthesis, the visual Turing test showed that even radiologists could not tell the GAN-synthesized from the raw images (accuracy: primary radiologist 56%, senior radiologist 65%). The experiments indicated that GAN augmentation method improved the classification accuracy by 23.5% (from 37.0% to 60.5%) and 7.3% (from 53.2% to 60.5%) in comparison with training methods using raw and classic augmented images respectively. Very similar results were also found with fine-tuning VGG16 under the same conditions, obtaining a classification accuracy of 37.7%, 48.3% and 60.2% for a training with the raw dataset, common augmentation dataset and GAN-synthesized dataset, respectively. Zebin and Rezvy [16] implemented a transfer learning pipeline for classifying COVID-19 chest x-ray images. The classifier effectively distinguishes inflammation in lungs due to COVID-19 and Pneumonia from the ones with no infection (normal). They have used multiple pre-trained (on ImageNet dataset) convolutional backbones as the feature extractor and achieved an overall detection accuracy of 88%, 94.3%, and 96.8% for VGG16, ResNet50, and EfficientNetB0 respectively when a basic data augmentation was employed. Additionally, they generated synthetic COVID-19 images with a CycleGAN to balance the three classes and then applied classic augmentation to all data. VGG16 model fine-tuned over this expanded database, produced an accuracy of 90%.

With regards to the second method, transfer learning has an incredible potential and can be fully applied when researchers have neither a sufficient volume of data nor the computational resources needed to train the algorithm. The resulting models will have an excellent features extraction capability learned from the large source datasets [31, 68]. However, they will be validated, tailored, and improved to the specific application to achieve optimal results. For brain tumor classification for MR images, Swati *et al* [2] used pre-trained deep CNN VGG-19 model, trained on ImageNet dataset, and proposed a block-wise fine-tuning strategy based on transfer learning, achieving the best average accuracy of 94.82% under five-fold cross-validation. They stated that thanks to transfer learning and fine-tuning it was possible to reduce overfitting and speed the convergence. Moreover, fine-tuning the last a few layers, it was be difficult for the CNN model to learn relevant medical brain MRI features from natural images. To achieve better performance, deep fine-tuning was required. As gradually increasing the layers for fine-tuning, the performance increased gradually. Hu *et al* [52] studied the diagnosis of prostate transition zone cancer (PTZC) versus benign prostatic hyperplasia on MRI. The deep CNN Alex-Net combined with transfer learning showed high efficacy in diagnosing PTZC on medical imaging, overcoming the challenge of limited data. Alex-Net was trained and compared between different transfer learning databases (ImageNet vs. disease-related images) and protocols (from scratch and fine-tuning). Using the model trained from scratch, authors obtained an AUC of 0.73. The efficacy of transfer learning from natural images was be limited (AUC of 0.75) but improved by transferring knowledge from the disease-related images (AUC of 0.86). Chougrad *et al* [51] aimed to classify mammography mass lesions as benign or malignant. To achieve this goal, they explored the importance of transfer learning and were able to fine-tune some of the most powerful CNNs (VGG16, ResNet50 and Inception v3, pre-trained on ImageNet). They also applied classic data augmentation and 5-fold cross validation during training. Due to the deep architectures and the small datasets used, they found that fione-tuning too many layers leads to worse results; the best fine-tuning strategy was to froze all the layers until the last or the two last convolutional blocks. The performance on the dataset used to fine-tune the model brought to a test accuracy of 98.64%, 98.77%, 98.94% for VGG16, ResNet50 and Inception v3, respectively. The best performing model was also tested on an independent database and got 98.23% accuracy.

Developing AI models that can learn from limited data is still an open research area, however these techniques not only tackle the insufficiency issue of data but can also provide a viable solution to class imbalance problem, which is also an important research area.

A central aspect that needs to be further explored is how the data augmentation affects the bias propagation. When the augmented data does not accurately reflect the real-world distribution, the model becomes biased. Bias refers to systematic errors or prejudices that exist in data, leading to unfair or discriminatory outcomes. When data augmentation techniques are applied, they can inadvertently amplify existing biases or introduce new biases into the augmented data. Data augmentation techniques modify the original data samples, potentially altering the distribution of the training data. Jain *et al* [96] in a recent study pointed out that, although one expects GANs to replicate the distribution of the original data, in real-world settings with limited data, finite training time and network capacity, the generated distribution can only capture a subset of the original distribution. In this scenario, GANs generate a distribution with significantly less diversity in one or several dimensions compared to the original data, bringing along the side-effect of amplifying the bias. The authors explored how the use of synthetic data generated by GANs, which are currently used in many different fields, are sensitive to this phenomenon. They analyzed how the societal biases, like gender and skin tone, present in a dataset of faces of engineering professors collected from a selection of U.S. Universities would be enhanced by using different types of GANs to generate synthetic data. The authors recommend a critical and conscious approach in the use of GANs for data augmentation. In fact, in some situations, even if the data might seem well balanced, they could be affected by some hidden bias and the augmented data might be under-representing some crucial feature of the real-world data. In those cases, the use of more reliable techniques should be considered.

Another important point that needs to be carefully investigated concerns the relationship between data augmentation and explainability. While data augmentation can significantly improve model performance by providing more varied and representative training examples, it can also have an impact on the explainability of machine learning models. Explainability refers to the ability to understand and interpret the decision-making process of a machine learning model. It is crucial in many domains where transparency, accountability, and trust are required, such as in healthcare. The impact of data augmentation on explainability can be examined from two perspectives: model interpretability and feature importance. In the first one, data augmentation can affect model interpretability by introducing additional complexity and non-linearity into the training process. When augmented data is used, the model is exposed to a wider range of input variations, making it more challenging to pinpoint the exact reasons for a particular decision. The transformations applied during augmentation can distort or alter the original features, making it harder to understand how the model is leveraging specific input characteristics to make predictions. In the second one, data augmentation can also influence feature importance analysis, which aims to identify the input features that have the most significant impact on the model predictions. By augmenting the data, the distribution and relationships between the features can change. This alteration can lead to changes in the perceived importance of certain features, as the model may rely more heavily on augmented features or combinations of features that were not present in the original dataset.

TRIPOD analysis brought out that most studies neither shared their source code nor included enough information about the model architecture, hyperparameters used, validation and evaluation methods followed to achieve such very good results. This leads to raising questions about the obtained results. Is not it that such exciting results were associated with some methodological bias that overestimates the performance of the resulting model? Moreover, limited accessibility of datasets and codes makes it difficult to assess the reproducibility of AI research. This approach is not constructive and affects external validity and denies implementation by other researchers that could improve the model. We strongly recommend more transparent reporting, sharing code, data (if possible) and detailing the hardware used. Only in this way can the replicability and robustness of the study be verified. Further, from the TRIPOD survey it emerged that it would be desirable to improve the drafting of the title and abstract by inserting more explanatory keywords.

Some limitations in our study can be highlighted. First, our search may have missed some studies that could have been included although comprehensive and systematic. Second, the guideline we used to assess the quality of the studies (TRIPOD) was not designed for AI studies, so some items and their adherence levels need some degree of interpretation. Third, we focused on studies that used small databases within clinical images; we believe it may not be appropriate to generalize our findings to other databases employed in the field of AI. Taking into account the main limitation emerged from this review, we feel compelled to underline the importance of the external validation of the developed models. This verification process aims to ensure the credibility, reliability, and accuracy of the results by subjecting them to scrutiny and evaluation by involving external, unbiased and independent validators. It helps mitigate biases and errors that might have been overlooked by the original researchers or developers. The external independent validation enhances the transparency and accountability of the research and development process and helps build trust among stakeholders, decision-makers, and the wider community. Overall, external validation is an important process for ensuring that models are performing as intended, and that the results are accurate and reliable against real-world data. In addition, it provides confidence in the decisions made based on the output of the model, essential in the clinical field.

As further suggestions for future directions, since data augmentation can impact bias propagation in machine learning models, caution must be exercised to ensure that biases are not amplified or introduced during the augmentation process. A thoughtful approach that includes diverse and representative data, bias detection and correction can help mitigate bias propagation. Furthermore, although data augmentation can pose challenges to model explainability, the following strategies can help mitigate these challenges: (i) careful consideration of methods specifically designed to improve the interpretability of models trained on augmented data, (ii) awareness of the impact of augmentation on feature importance, and (iii) controlled augmentation strategies to ensure that the augmented data samples preserve the salient characteristics of the original data. In our opinion this topic is not explicitly addressed in the literature and it should be developed in future works. Ultimately, balancing the benefits of improved model performance with the need for interpretability is essential, particularly in domains where transparency and accountability are critical. For this purpose, post-hoc interpretability methods should be employed by highlighting relevant features or generating saliency maps.

5. Qualitative summary

The research question of this systematic review is to highlight the different strategies for working with small data. On the basis of the research question, the following qualitative summary may be extracted from the surveyed papers. Working with AI in a small medical database requires careful consideration of various strategies to ensure effective utilization of the available data and the AI capabilities. First of all, ensuring the quality of your data is paramount and cleaning and preprocessing the data to remove noise, errors, and inconsistencies will improve the accuracy of any AI models you develop. Moreover, with limited data to make the most out of the available information, feature engineering and selection becomes particularly important by identifying the most relevant features. As for algorithms selection and training, the combination of regularization to penalize complex models and k-fold cross-validation to assess the generalization performance on unseen data could mitigate the overfitting with limited data. As emerged from the analysis of the examined publications the two most widely used approaches to cope with the issues in the use of the small databases are data augmentation and transfer learning, also thanks to their simple application and diffusion reflecting their effectiveness and versatility across various domains. The first one (i) supports to introduce variability into the training data, making AI models more robust to variations and noise present in real-world scenarios, (ii) encourages the model to learn more invariant and discriminative features, improving its generalization performance and (iii) helps prevent the model from focusing too heavily on idiosyncratic patterns in the training data, reducing the risk of overfitting. The second one (i) by initializing the model with weights learned from a pre-trained model and fine-tuning it on the target dataset, enables faster convergence and often results in better performance compared to training from scratch, (ii) can facilitate domain adaptation, where knowledge learned from a source domain is adapted to a target domain with different characteristics and (iii) can achieve state-of-the-art performance without the need for extensive computational resources or labeled data.

The analysis revealed that limiting the definition of ‘small’ to a database considering only the number of samples or records it contains is inadequate. Hence, it becomes imperative to recalibrate the understanding of scale, shifting the focus from absolute metrics to a more nuanced perspective. Instead of merely counting entries, the intricate interplay between data volume and the number of free parameters awaiting optimization must be considered. By contextualizing the term ‘small’ within the framework of relative proportions, it becomes apparent that the size of the database alone does not dictate the complexity of the task. Rather, it is the ratio between the volume of data and the degree of freedom within the model that truly defines the magnitude of the challenge. This is the fundamental aspect, but the analysis also outlined other common characteristics of the small databases. Specifically, they may lack diversity in terms of the range of instances or scenarios they cover, have a limited number of features or attributes for each sample, may suffer from imbalanced classes and may contain more noise or variability compared to larger datasets.

When dealing with machine learning models trained on small datasets, a key concern is the risk of overfitting, which may restrict their ability to generalize beyond the training data. However, employing an explainable artificial intelligence based solution ensures that those assessing the model can acquire the necessary insights to conduct specific evaluations of its reliability and effectiveness [97]. Specifically, by tracking how the importance of features varies across different data segments, it becomes possible to gauge whether the factors driving model decisions are changing.

By incorporating the aforementioned strategies into AI workflows, it is possible to mitigate the challenges associated with small medical databases and develop robust and accurate AI models across a wide range of applications and domains.

6. Conclusions

Though AI requires a sufficient amount of quality data for training, the results obtained using small databases of medical images are promising but still not mature enough to be implemented in the clinical setting and be widely used. Transfer learning and data augmentation could represent the most reasonable choices to fight overfitting. Despite the good performances obtained so far, often too promising, there is still a lot of work to be done. First of all, to encourage the external validation of the models, using databases that are independent from those of the training. Consequently, it is necessary to sensitize researchers to be more transparent, sharing codes and data as much as possible. This attitude will help the reproducibility, the generalizability and the development of higher quality research.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Conflict of interest

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

Funding

NextAIM project by National Institute for Nuclear Physic funded this study.

ORCID iDs

Stefano Piffer https://orcid.org/0000-0002-6474-2885
Sabina Tangaro https://orcid.org/0000-0002-1372-3916
Alessandra Retico https://orcid.org/0000-0001-5135-4472
Cinzia Talamonti https://orcid.org/0000-0003-2955-6451

References

[1] Romero M, Interian Y, Solberg T and Valdes G 2020 Targeted transfer learning to improve performance in small medical physics datasets *Med. Phys.* **47** 6246–56

[2] Swati Z N K, Zhao Q, Kabir M, Ali F, Ali Z, Ahmed S and Lu J 2019 Brain tumor classification for MR images using transfer learning and fine-tuning *Comput. Med. Imaging Graph.* **75** 34–46

[3] D'souza R N, Huang P-Y and Yeh F-C 2020 structural analysis and optimization of convolutional neural networks with a small sample size *Sci. Rep.* **10** 834

[4] Ubaldi L *et al* 2021 Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples *Phys. Med.* **90** 13–22

[5] Vabalas A, Gowen E, Poliakoff E and Casson A J 2019 Machine learning algorithm validation with a limited sample size *PLoS One* **14** 1–20

[6] Xu Y *et al* 2021 A hierarchical deep learning approach with transparency and interpretability based on small samples for glaucoma diagnosis *npj Digit. Med.* **4** 1–11

[7] Ayana G, Park J, Jeong J-W and Choe S 2022 A novel multistage transfer learning for ultrasound breast cancer image classification *Diagnostics* **12** 135

[8] Trivizakis E, Manikis G C, Nikiforaki K, Drevelegas K, Constantinides M, Drevelegas A and Marias K 2019 Extending 2-D convolutional neural networks to 3-D for advancing deep learning cancer classification with application to MRI liver tumor differentiation *IEEE J. Biomed. Health Inform.* **23** 923–30

[9] Han S, Williamson B D and Fong Y 2021 Improving random forest predictions in small datasets from two-phase sampling designs *BMC Med. Inf. Decis. Mak.* **21** 1–9

[10] Adedigba A P, Adeshina S A and Aibinu A M 2022 Performance evaluation of deep learning models on mammogram classification using small dataset *Bioengineering* **9** 161

[11] Gatidis S *et al* 2015 Combined unsupervised-supervised classification of multiparametric PET/MRI data: application to prostate cancer *NMR Biomed.* **28** 914–22

[12] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial networks (arXiv:1406.2661)

[13] Pan T, Chen J, Zhang T, Liu S, He S and Lv H 2022 Generative adversarial network in mechanical fault diagnosis under small sample: a systematic review on applications and future perspectives *ISA Trans.* **128** 1–10

[14] Levine A B *et al* 2020 Synthesis of diagnostic quality cancer pathology images by generative adversarial networks *J. Pathol.* **252** 178–88

[15] Shi G, Wang J, Qiang Y, Yang X, Zhao J, Hao R, Yang W, Du Q and Kazihise N G-F 2020 Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification *Comput. Methods Programs Biomed.* **196** 105611

[16] Zebin T and Rezvy S 2021 COVID-19 detection and disease progression visualization: deep learning on chest x-rays for classification and coarse localization *Appl. Intell.* **51** 1010–21

[17] Gheshlaghi S H, Nok Enoch Kan C and Ye D H 2021 Breast cancer histopathological image classification with adversarial image synthesis *2021 43rd Annual Int. Conf. IEEE Engineering in Medicine and Biology Society* (IEEE) pp 3387–90

[18] Alruwaili M and Gouda W 2022 Automated breast cancer detection models based on transfer learning *Sensors* **22** 876

[19] Bahgat W M, Balaha H M, AbdulAzeem Y and Badawy M M 2021 An optimized transfer learning-based approach for automatic diagnosis of COVID-19 from chest x-ray images *PeerJ. Comput. Sci.* **7** 1–14

[20] Horry M J, Chakraborty S, Paul M, Ulhaq A, Pradhan B, Saha M and Shukla N 2020 COVID-19 detection through transfer learning using multimodal imaging data *IEEE Access* **8** 149808–24

[21] Russakovsky O *et al* 2015 ImageNet large scale visual recognition challenge *Int. J. Comput. Vis.* **115** 211–52

[22] Aderghal K, Afdel K, Benois-Pineau J and Catheline G 2020 Improving Alzheimer's stage categorization with Convolutional Neural Network using transfer learning and different magnetic resonance imaging modalities *Heliyon* **6** e05652

[23] Sanchez K, Hinojosa C, Arguello H, Kouame D, Meyrignac O and Basarab A 2022 CX-DaGAN: domain adaptation for pneumonia diagnosis on a small chest x-ray dataset *IEEE Trans. Med. Imaging* **41** 3278–88

[24] Sha X, Gong G, Qiu Q, Duan J, Li D and Yin Y 2019 Identifying pathological subtypes of non-small-cell lung cancer by using the radiomic features of 18F-fluorodeoxyglucose positron emission computed tomography *Transl. Cancer Res.* **8** 1741–9

[25] Moher D, Liberati A, Tetzlaff J and Altman D G 2009 Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement *BMJ* **339** 332–6

[26] Collins G S, Reitsma J B, Altman D G and Moons K G M 2015 Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement *BMJ* **350** 1–9

[27] Moons K G M, Altman D G, Reitsma J B, Ioannidis J P A, Macaskill P, Steyerberg E W, Vickers A J, Ransohoff D F and Collins G S 2015 Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration *Ann. Intern. Med.* **162** W1–W73

[28] Heus P, Damen J A A G, Pajouheshnia R, Scholten R J P M, Reitsma J B, Collins G S, Altman D G, Moons K G M and Hooft L 2019 Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies *BMJ Open* **9** 1–6

[29] Abbasi S, Hajabdollahi M, Khadivi P, Karimi N, Roshandel R, Shirani S and Samavi S 2021 Classification of diabetic retinopathy using unlabeled data and knowledge distillation *Artif. Intell. Med.* **121** 102176

[30] Ali Khan H, Jue W, Mushtaq M and Umer Mushtaq M 2020 Brain tumor classification in MRI image using convolutional neural network *Math. Biosci. Eng.* **17** 6203–16

[31] Alzubaidi L, Al-Amidie M, Al-Asadi A, Humaidi A J, Al-Shamma O, Fadhel M A, Zhang J, Santamaría J and Duan Y 2021 Novel transfer learning approach for medical imaging with limited labeled data *Cancers* **13** 1590

[32] Wodzinski M, Banzato T, Atzori M, Andrearczyk V, Cid Y D and Muller H 2020 Training deep neural networks for small and highly heterogeneous MRI datasets for cancer grading *2020 42nd Annual Int. Conf. IEEE Engineering in Medicine and Biology Society* (IEEE) pp 1758–61

[33] Hertel R and Benlamri R 2021 COV-SNET: a deep learning model for x-ray-based COVID-19 classification *Inform. Med. Unlocked.* **24** 100620

[34] Baydilli Y Y and Atila Ü 2020 Classification of white blood cells using capsule networks *Comput. Med. Imaging Graph.* **80** 101699

[35] Li F, Chen H, Liu Z, Zhang X, Jiang M, Wu Z and Zhou K 2019 Deep learning-based automated detection of retinal diseases using optical coherence tomography images *Biomed. Opt. Express* **10** 6204

[36] Xia T, Kumar A, Feng D and Kim J 2018 Patch-level tumor classification in digital histopathology images with domain adapted deep learning *40th Annual Int. Conf. IEEE IEEE Engineering in Medicine and Biology Society* (https://doi.org/10.1109/EMBC.2018.8512353)

[37] Shen Y, Li X, Liang X, Xu H, Li C, Yu Y and Qiu B 2020 A deep-learning-based approach for adenoid hypertrophy diagnosis *Med. Phys.* **47** 2171–81

[38] Feng Y, Zhang L and Yi Z 2018 Breast cancer cell nuclei classification in histopathology images using deep neural networks *Int. J. Comput. Assist. Radiol. Surg.* **13** 179–91

[39] Liu S, Shah Z, Sav A, Russo C, Berkovsky S, Qian Y, Coiera E and Di Ieva A 2020 Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning *Sci. Rep.* **10** 1–11

[40] Ahn E, Kumar A, Fulham M, Feng D and Kim J 2020 Unsupervised domain adaptation to classify medical images using zero-bias convolutional auto-encoders and context-based feature augmentation *IEEE Trans. Med. Imaging* **39** 2385–94

[41] Montoya J C, Li Y, Strother C and Chen G H 2018 3D deep learning angiography (3D-DLA) from C-arm conebeam CT *Am. J. Neuroradiol.* **39** 916–22

[42] Xia X, Gong J, Hao W, Yang T, Lin Y, Wang S and Peng W 2020 Comparison and fusion of deep learning and radiomics features of ground-glass nodules to predict the invasiveness risk of stage-I lung adenocarcinomas in CT scan *Front. Oncol.* **10** 418

[43] Liang G, Xing X, Liu L, Zhang Y, Ying Q, Lin A-L and Jacobs N 2021 Alzheimer's disease classification using 2D convolutional neural networks *2021 43rd Annual Int. Conf. IEEE Engineering in Medicine and Biology Society* (IEEE) pp 3008–12

[44] Huynh B Q, Li H and Giger M L 2016 Digital mammographic tumor classification using transfer learning from deep convolutional neural networks *J. Med. Imaging* **3** 034501

[45] Zhang X, Yang Y, Li T, Zhang Y, Wang H and Fujita H 2021 CMC: a consensus multi-view clustering model for predicting Alzheimer's disease progression *Comput. Methods Programs Biomed.* **199** 105895

[46] Hu M, Sim K, Zhou J H, Jiang X and Guan C 2020 Brain MRI-based 3D convolutional neural networks for classification of schizophrenia and controls *2020 42nd Annual Int. Conf. IEEE Engineering in Medicine and Biology Society* (IEEE) pp 1742–5

[47] Dai Y, Gao Y and Liu F 2021 TransMed: transformers advance multi-modal medical image classification *Diagnostics* **11** 1384

[48] Apostolopoulos I D, Papathanasiou N D, Spyridonidis T and Apostolopoulos D J 2020 Automatic characterization of myocardial perfusion imaging polar maps employing deep learning and data augmentation *Hell. J. Nucl. Med.* **23** 125–32

[49] Bien N *et al* 2018 Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet *PLOS Med.* **15** e1002699

[50] Fu Y, Xue P, Ji H, Cui W and Dong E 2020 Deep model with Siamese network for viable and necrotic tumor regions assessment in osteosarcoma *Med. Phys.* **47** 4895–905

[51] Chougrad H, Zouaki H and Alheyane O 2018 Deep convolutional neural networks for breast cancer screening *Comput. Methods Programs Biomed.* **157** 19–30

[52] Hu B *et al* 2021 Classification of prostate transitional zone cancer and hyperplasia using deep transfer learning from disease-related images *Cureus* **13** e14108

[53] Ye J, Luo J, Xu S and Wu W 2020 One-slice CT image based kernelized radiomics model for the prediction of low/mid-grade and high-grade HNSCC *Comput. Med. Imaging Graph.* **80** 101675

[54] Zhou J *et al* 2021 Automatic detection and classification of focal liver lesions based on deep convolutional neural networks: a preliminary study *Front. Oncol.* **10** 1–11

[55] Yi P H, Lin A, Wei J, Yu A C, Sair H I, Hui F K, Hager G D and Harvey S C 2019 Deep-learning-based semantic labeling for 2D mammography and comparison of complexity for machine learning tasks *J. Digit. Imaging* **32** 565–70

[56] Mutasa S, Chang P D, Ruzal-Shapiro C and Ayyala R 2018 MABAL: a novel deep-learning architecture for machine-assisted bone age labeling *J. Digit. Imaging* **31** 513–9

[57] Mzoughi H, Njeh I, Wali A, Ben Slima M, BenHamida A, Mhiri C and Ben Mahfoudhe K 2020 Deep multi-scale 3D convolutional neural network (CNN) for MRI gliomas brain tumor classification *J. Digit. Imaging* **33** 903–15

[58] Wang C, Elazab A, Wu J and Hu Q 2017 Lung nodule classification using deep feature fusion in chest radiography *Comput. Med. Imaging Graph.* **57** 10–18

[59] Samala R K, Chan H-P, Hadjiiski L M, Helvie M A and Richter C D 2020 Generalization error analysis for deep convolutional neural network with transfer learning in breast cancer diagnosis *Phys. Med. Biol* **65** 105002

[60] Yi P H, Kim T K, Wei J, Shin J, Hui F K, Sair H I, Hager G D and Fritz J 2019 Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning *Pediatr. Radiol.* **49** 1066–70

[61] An G, Akiba M, Omodaka K, Nakazawa T and Yokota H 2021 Hierarchical deep learning models using transfer learning for disease detection and classification based on small number of medical images *Sci. Rep.* **11** 4250

[62] Owais M, Arsalan M, Mahmood T, Kang J K and Park K R 2020 Automated diagnosis of various gastrointestinal lesions using a deep learning–based classification and retrieval framework with a large endoscopic database: model development and validation *J. Med. Internet Res.* **22** e18563

[63] Samala R K, Chan H P, Hadjiiski L and Helvie M A 2021 Risks of feature leakage and sample size dependencies in deep feature extraction for breast mass classification *Med. Phys.* **48** 2827–37

[64] Cogan T, Cogan M and Tamil L 2019 MAPGI: accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning *Comput. Biol. Med.* **111** 103351

[65] Apostolopoulos I D, Aznaouridis S I and Tzani M A 2020 Extracting possibly representative COVID-19 biomarkers from x-ray images with deep learning approach and image data related to pulmonary diseases *J. Med. Biol. Eng.* **40** 462–9

[66] Choi J Y, Yoo T K, Seo J G, Kwak J, Um T T and Rim T H 2017 Multi-categorical deep learning neural network to classify retinal images: a pilot study employing small database *PLoS One* **12** e0187336

[67] Zong W *et al* 2020 A deep dive into understanding tumor foci classification using multiparametric MRI based on convolutional neural network *Med. Phys.* **47** 4077–86

[68] Uemura T, Näppi J J, Ryu Y, Watari C, Kamiya T and Yoshida H 2021 A generative flow-based model for volumetric data augmentation in 3D deep learning for computed tomographic colonography *Int. J. Comput. Assist. Radiol. Surg.* **16** 81–89

[69] Oakden-Rayner L, Carneiro G, Bessen T, Nascimento J C, Bradley A P and Palmer L J 2017 Precision Radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework *Sci. Rep* **7** 1648

[70] Nabizadeh-Shahre-Babak Z, Karimi N, Khadivi P, Roshandel R, Emami A and Samavi S 2021 Detection of COVID-19 in x-ray images by classification of bag of visual words using neural networks *Biomed. Signal Process. Control* **68** 102750

[71] Wang Y, Zhou L, Wang M, Shao C, Shi L, Yang S, Zhang Z, Feng M, Shan F and Liu L 2020 Combination of generative adversarial network and convolutional neural network for automatic subcentimeter pulmonary adenocarcinoma classification *Quant. Imaging Med. Surg.* **10** 1249–64

[72] Haga A, Takahashi W, Aoki S, Nawa K, Yamashita H, Abe O and Nakagawa K 2018 Classification of early stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: interobserver delineation variability analysis *Radiol. Phys. Technol.* **11** 27–35

[73] Fantini I, Yasuda C, Bento M, Rittner L, Cendes F and Lotufo R 2021 Automatic MR image quality evaluation using a Deep CNN: a reference-free method to rate motion artifacts in neuroimaging *Comput. Med. Imaging Graph.* **90** 101897

[74] Zhang S, Han F, Liang Z, Tan J, Cao W, Gao Y, Pomeroy M, Ng K and Hou W 2019 An investigation of CNN models for differentiating malignant from benign lesions using small pathologically proven datasets *Comput. Med. Imaging Graph.* **77** 101645

[75] Toda R, Teramoto A, Tsujimoto M, Toyama H, Imaizumi K, Saito K and Fujita H 2021 Synthetic CT image generation of shape-controlled lung cancer using semi-conditional InfoGAN and its applicability for type classification *Int. J. Comput. Assist. Radiol. Surg.* **16** 241–51

[76] Usman M, Zia T and Tariq A 2022 Analyzing transfer learning of vision transformers for interpreting chest radiography *J. Digit. Imaging* **35** 1445–62

[77] Kaur T and Gandhi T K 2022 Classifier fusion for detection of COVID-19 from CT scans *Circuits Syst. Signal Process.* **41** 3397–414

[78] Hashemzehi R, Seyyed Mahdavi S J, Kheirabadi M and Kamel S R 2021 Y-net: a reducing gaussian noise convolutional neural network for MRI brain tumor classification with NADE concatenation *Biomed. Phys. Eng. Express* **7** 055006

[79] Rocca A *et al* 2021 Early diagnosis of liver metastases from colorectal cancer through CT radiomics and formal methods: a pilot study *J. Clin. Med.* **11** 31

[80] Suganyadevi S and Seethalakshmi V 2022 CVD-HNet: classifying pneumonia and COVID-19 in chest x-ray images using deep network *Wirel. Pers. Commun.* **126** 3279–303

[81] Ahmad B, Sun J, You Q, Palade V and Mao Z 2022 Brain tumor classification using a combination of variational autoencoders and generative adversarial networks *Biomedicines* **10** 223

[82] Le W T, Vorontsov E, Romero F P, Seddik L, Elsharief M M, Nguyen-Tan P F, Roberge D, Bahig H and Kadoury S 2022 Cross-institutional outcome prediction for head and neck cancer patients using self-attention neural networks *Sci. Rep.* **12** 3183

[83] Cahan N, Marom E M, Soffer S, Barash Y, Konen E, Klang E and Greenspan H 2022 Weakly supervised attention model for RV strain classification from volumetric CTPA scans *Comput. Methods Programs Biomed.* **220** 106815

[84] Ho T K K and Gwak J 2022 Feature-level ensemble approach for COVID-19 detection using chest x-ray images *PLoS One* **17** e0268430

[85] Muhammad K *et al* 2022 WEENet: an intelligent system for diagnosing COVID-19 and lung cancer in IoMT environments *Front. Oncol.* **11** 1–13

[86] Sahoo P, Roy I, Ahlawat R, Irtiza S and Khan L 2022 Potential diagnosis of COVID-19 from chest x-ray and CT findings using semi-supervised learning *Phys. Eng. Sci. Med.* **45** 31–42

[87] Zhang Z, Ji Z, Chen Q, Yuan S and Fan W 2022 Joint optimization of CycleGAN and CNN classifier for detection and localization of retinal pathologies on color fundus photographs *IEEE J. Biomed. Health Inform.* **26** 115–26

[88] Ben Ahmed K, Hall L O, Goldgof D B and Gatenby R 2022 Ensembles of convolutional neural networks for survival time estimation of high-grade glioma patients from multimodal MRI *Diagnostics* **12** 345

[89] Ettehadi N, Kashyap P, Zhang X, Wang Y, Semanek D, Desai K, Guo J, Posner J and Laine A F 2022 Automated multiclass artifact detection in diffusion MRI volumes via 3D residual squeeze-and-excitation convolutional neural networks *Front. Hum. Neurosci.* **16** 877326

[90] DeVries T and Taylor G W 2017 Improved regularization of convolutional neural networks with cutout (arXiv:1708.04552)

[91] Homeyer A *et al* 2022 Recommendations on compiling test datasets for evaluating artificial intelligence solutions in pathology *Mod. Pathol.* **35** 1759–69

[92] Yu A C, Mohajer B and Eng J 2022 External validation of deep learning algorithms for radiologic diagnosis: a systematic review *Radiol. Artif. Intell.* **4** 1–9

[93] Heim E, Roß T, Seitel A, März K, Stieltjes B, Eisenmann M, Lebert J, Metzger J and Sommer G 2018 Large-scale medical image annotation with crowd-powered algorithms *J. Med. Imaging* **5** 1

[94] Navarro F *et al* 2021 Development and external validation of deep-learning-based tumor grading models in soft-tissue sarcoma patients using MR imaging *Cancers* **13** 2866

[95] Sarica B and Seker D Z 2022 New MS lesion segmentation with deep residual attention gate U-Net utilizing 2D slices of 3D MR images *Front. Neurosci.* **16** 912000

[96] Jain N, Olmo A, Sengupta S, Manikonda L and Kambhampati S 2022 Imperfect ImaGANation: implications of GANs exacerbating biases on facial data augmentation and snapchat face lenses *Artif. Intell.* **304** 103652

[97] Lombardi A, Diacono D, Amoroso N, Monaco A, Tavares J M R S, Bellotti R and Tangaro S 2021 Explainable deep learning for personalized age prediction with brain morphology *Front. Neurosci.* **15** 674055