UNIVERSITÀ
DEGLI STUDI
FIRENZE

PHD PROGRAM IN SMART COMPUTING
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE (DINFO)

Dealing with Small Datasets in Artificial Intelligence: focus on Medical Imaging

Stefano Piffer

Dissertation presented in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Smart Computing

PhD Program in Smart Computing
University of Florence, University of Pisa, University of Siena

Dealing with Small Datasets in Artificial Intelligence: focus on Medical Imaging

Stefano Piffer

Advisor:

Prof. Cinzia Talamonti

Head of the PhD Program:

Prof. Stefano Berretti

Evaluation Committee:

Prof. Sabina Sonia Tangaro,

Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli studi di Bari Aldo Moro, Bari, Italy

Prof. Maurizio Marrale,

Dipartimento di Fisica e Chimica, Università di Palermo, Palermo, Italy

XXXV ciclo — July 2023

Acknowledgments

I would first like to thank my Ph.D. supervisor Prof. Cinzia Talamonti for her guidance, support and overseeing during this long journey. Also, I would like to express my gratitude to Dr. Andrea Barucci and Dr. Lucio Anderlini for sharing their expertise and for the advices they gave me, and to the reviewers Prof. Sonia Tangaro and Prof. Maurizio Marrale.
This work has been carried out within the Artificial Intelligence in Medicine (AIM) CSN5, 2019 − 2021) and next AIM (CSN5, 2022 − 2024) projects funded by National Institute of Nuclear Physics (INFN).
A very profound gratitude goes to Francesca, for providing me with unfailing support and continuous encouragement throughout these years and for listening to me while I bore her about the research.

Abstract

Artificial Intelligence (AI) will likely affect healthcare systems significantly and it could play a key role in clinical decision making in future. Deep learning and radiomics methods are extremely promising machine learning tools to analyze complex and high-dimensional medical images. But unfortunately, machine learning models that work with imaging data require massive amounts of data. For this reason, their implementation in healthcare settings remains limited, mostly due to the lack of very large datasets on which to test the generalizability and reliability of the trained models. Although many institutes are collaborating to produce publicly available datasets of medical images, the access to medical images remains limited and the small sample sizes and lack of diverse geographic areas hinder the generalizability and accuracy of developed solutions. Moreover, the process of data acquisition is severely limited by different challenges. These obstacles are mainly related to privacy regulations and the effort of domain experts to assess imaging data quality and produce high-quality ground truth. Medical data are often stored in disparate silos which in turn results in the difficulty of managing large medical imaging datasets. Furthermore, simply achieving access to large quantities of image data is insufficient to allay these shortcomings. Adequate curation, analysis, labeling, and clinical application are critical to achieving high-impact clinically meaningful AI algorithms.

This Ph.D. thesis describes the process of labeling, curating, managing and sharing medical image data for AI algorithm development for optimal clinical impact, while maintaining a high degree of privacy and security in exchanging sensitive data. The pros and cons of having heterogeneous or homogeneous data have been taken into consideration. The first, caused by the diversity of the populations included in the dataset, leads to incompleteness for the different data acquisition standards and practices. The second, although it returns complete and uniform datasets, does not fully consider the natural variability of the population. This work provides an application of the various approaches proposed in the literature to alleviate the problem of small data samples in AI. Well-established techniques such as unsupervised hierarchical clustering and transfer learning in the context of rare diseases stratification have been analyzed. Moreover, a U-net was trained from scratch with the help of data augmentation merging public datasets while trying to contain data and label heterogeneity.

The results are promising, showing that transfer learning technique can enable the training of custom models on small datasets by exploiting the powerful feature extraction modules of Convolutional Neural Networks. Different methods to select and combine features allow to incorporate more information and to reach high level of abstraction which in our case led to a natural clustering of data. Moreover, data augmentations combing different public dataset is also an effective technique to carry out a complete training.

In clinical context, build effective models based on small data is an urgent task since machine learning systems allow the identification of extremely difficult correlations among medical imaging and clinical endpoint. This path is viable, there are the right tools to deal with it, but one need to know how to use them with full knowledge of the facts, adapting them to the needs of the case.

This work has been developed in the framework of the INFN-funded AIM projects, that aims to exploit the expertise of INFN and associated researchers on medical data processing and enhancement, and turn it in the development of advanced and effective analysis instruments to be eventually clinically validated.

Contents

List of Figures

List of Tables

List of Acronyms

AIM Artificial Intelligence in Medicine

INFN National Institute of Nuclear Physics

AI Artificial Intelligence

ML Machine Learning

DL Deep Learning

CT Computed Tomography

MRI Magnetic Resonance Imaging

CV Cross-Validation

nCV nested Cross-Validation

TL Transfer Learning

NN Neural Network

DNN Deep Neural Network

CNN Convolutional Neural Network

GAN Generative Adversarial Network

MR Magnetic Resonance

DICOM Digital Imaging and Communications in Medicine

PACS Picture Archiving and Communication System

ROI Region Of Interest

NIfTI Neuroimaging Informatics Technology Initiative

MB Medulloblastoma

CSI Cranio-Spinal Irradiation

RFE Recursive Feature Elimination

RF Random Forest

XGB Extreme Gradient Boosting

HC Hierarchical Clustering

LG Logistic Regression

DT Decision Tree

GB Gradient Boosting

MCC Matthews Correlation Coefficient

RT Radiotherapy

MI Mutual Information

PCA Principal Component Analysis

WHO World Health Organization

PHI Protected Health Information

ICU Intensive Care Unit

FOV Field Of View

DSC Dice Similarity Coefficient

sDSC surface Dice Similarity Coefficient

vDSC volumetric Dice Similarity Coefficient

GGO Ground-Glass Opacifications

CT-SS CT-Severity Score

MAE Mean Absolute Error

STS Soft-Tissue-Sarcoma

CE Contrast-Enhanced

Grad-CAM Gradient Weighted Class Activation Map

NSCLC Non-Small Cell Lung Cancer

TCIA The Cancer Imaging Archive

PET Positron Emission Tomography

SPECT Single Photon Emission Computed Tomography

H&N Head & Neck

PRISMA Preferred Reporting Items for Systematic reviews and Meta-Analyses

TRIPOD Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Chapter 1

Introduction

AI, described as being able to perform tasks that normally require human-like cognitive functions, is having considerable potential in all areas of medicine (Langlotz et al., 2019; Hashimoto et al., 2018; Shen et al., 2020a), exponentially rising in popularity over recent years (Cai et al., 2020). With the advent of precision and personalized medicine, AI has received great interest as a promising tool for identifying the best diagnosis and treatment for an individual patient.

The aim is to achieve patient stratifications, monitoring and treatment design using quantitative, patient-specific datasets, integrated via algorithmic analyses. This implies embedding diagnostics and treatments with features derived from an advanced data analysis in order to create complete datasets describing multivariate aspects of individuals' health across time. Thus, it is also essential to identify measurable and accurate indicators, the biomarkers, which potentially can predict disease initiation and progression. To exploit the potential of such datasets, it is necessary to develop transparent and replicable mathematical frameworks able to describe and/or extract information from high-dimensional, dynamic, noisy and sparsely sampled processes to highlight time patterns in a disease. For this reason, we need mathematical modeling methods and statistical data analysis algorithms to be robust and able to adapt to errors and uncertainties.

Mathematical modeling can be mechanistic and non-mechanistic, such as AI techniques. The mechanistic models focus on the description of elements forming a system, their mutual interactions and the interaction with the environment with the possibility to also describe the resulting emerging behavior and average properties of the systems. AI models, instead, aim to simulate the logical decision-making process taking advantage of available data. These models can predict the behavior of a system searching for relationships between inputs and outputs or identifying specific or recurrent patterns. In other words, in classical science rules are applied to

data to obtain results, while in AI data and results are fed to algorithms to obtain rules.

Both Machine Learning (ML) and Deep Learning (DL) are subsections of artificial intelligence. Machine learning is the discipline that builds mathematical models by recognizing common patterns and uncovering disease characteristics through the use of a large number of handcrafted features manually extracted from data (Sollini et al., 2019) without being explicitly programmed to conduct these tasks (Avanzo et al., 2020). Deep learning algorithms can be seen as a combination of simple non-linear functions with the potential to model very complex systems, capable to automatically learn a set of features, usually over a certain number of layers making the model 'deep', from a labelled dataset and compute the final desired task (Yamashita et al., 2018).

Machine learning and deep learning algorithms have shown great potential in streamlining clinical task such as improve workflows in healthcare systems or to assist clinicians in decision making by automating tasks such as lesion detection or medical imaging quantification. The first category includes prioritizing worklists (JL, 2017) and triaging (Yala et al., 2019), while the second one covers detection of pulmonary nodules (Hwang et al., 2019), Alzheimer disease (Ding et al., 2019), and urinary stones (Parakh et al., 2019), CT coronary calcium scoring (de Vos et al., 2019), MRI prostate classification (Schelb et al., 2019), mammography breast density (Lehman et al., 2019).

To train such models, a great amount of available data is fundamental. Just think that in a radiomics study, a ML approach applied to the quantitative analysis of radiological images, up to some hundreds or thousands of features can be reached and in DL a network can manage millions of parameters to be optimized (Avanzo et al., 2021). This causes that, nearly all limitations can be attributed to one substantial problem: lack of available image data for training and testing of AI algorithms. Currently, most research groups have limited access to medical images, while the small sample sizes and lack of diverse geographic areas hinder the generalizability and accuracy of developed solutions (Soffer et al., 2019).

Although small data sets may be sufficient for training of AI algorithms in the research setting, large data sets with high-quality images and annotations are still essential for supervised training, validation, and testing of commercial AI algorithms. This is especially true in the clinical setting (Park and Han, 2018). In my short career I have had the pleasure to use only a clinically validated system available on the market, which performs both chest structures segmentation and diseases classification on medical images. A key aspect for the success of these deep learning models on these tasks is the availability of large labeled datasets of medical images,

containing thousands of examples (more than 10000) and a continuous reinforcement learning. However, it is challenging to curate large labeled medical imaging datasets of that scale and indeed such a database has been wanted and created following the $COVID-19$ pandemic. In recent years, some international competitions have released rich labelled medical images, which provides a potential data source to train models specific to medical applications (`https://grand-challenge.org`).

Taking up and deepening the discussion mentioned above, ML and DL algorithms require a large amount of training samples and an unappropriated sample size will lead to a reduction in the confidence of the prediction. As a matter of fact, datasets used for training AI have a small number of samples with respect to the dimensionality of data and of the desired tasks (Torgyn et al., 2015), to the point that, frequently, there are more features per subject than subjects in the entire dataset (Chatterjee et al., 2018). Under these circumstances, overfitting, a condition where models are more sensitive to noise in the data than to their patterns, and instability occur, making the model poorly reproducible and generalizable, meaning that it will perform badly on unseen datasets (Cui et al., 2020; Nensa et al., 2019). In other words, overfitting arises when the algorithm fits the training data too tightly; this leads to overestimated results and the model loses the ability to generalize on new data. However, techniques and methodologies have been developed to minimize this pitfall, such as following appropriate data pre-processing procedures (features selection (Parmar et al., 2015; Lian et al., 2016)) or implementing cutting-edge AI algorithms (Generative Adversarial Network (Zhang et al., 2022a; Talha and Hazrat, 2018; Ma et al., 2021b)).

The easiest way to try to compensate for overfitting is Cross-Validation (CV). K-fold cross-validation is the most common technique for model validation and model selection (Rodriguez et al., 2010), based on the idea that each sample in the dataset has the opportunity of being a test sample. The process involves splitting the dataset into k parts and the model is trained k times. Each time one part is used as a test set and the other $k-1$ parts are merged and used to train the model (Figure 1.1). By doing this, each sample in the dataset will get a chance to be a test sample. The main drawback of this approach is that, since the model selection is performed on the whole dataset, split into k folds, there is no separate test set to estimate the chosen best model's generalization ability.

To overcome this drawback, another more robust method of model validation, namely nested Cross-Validation (nCV), is recommended in most AI applications for small to moderate-sized datasets. nCV, which was first described by Varma and Simon (Rivals et al., 2007) when working with small datasets, is a procedure that offers a workaround for small-dataset situations that shows a low bias in practices

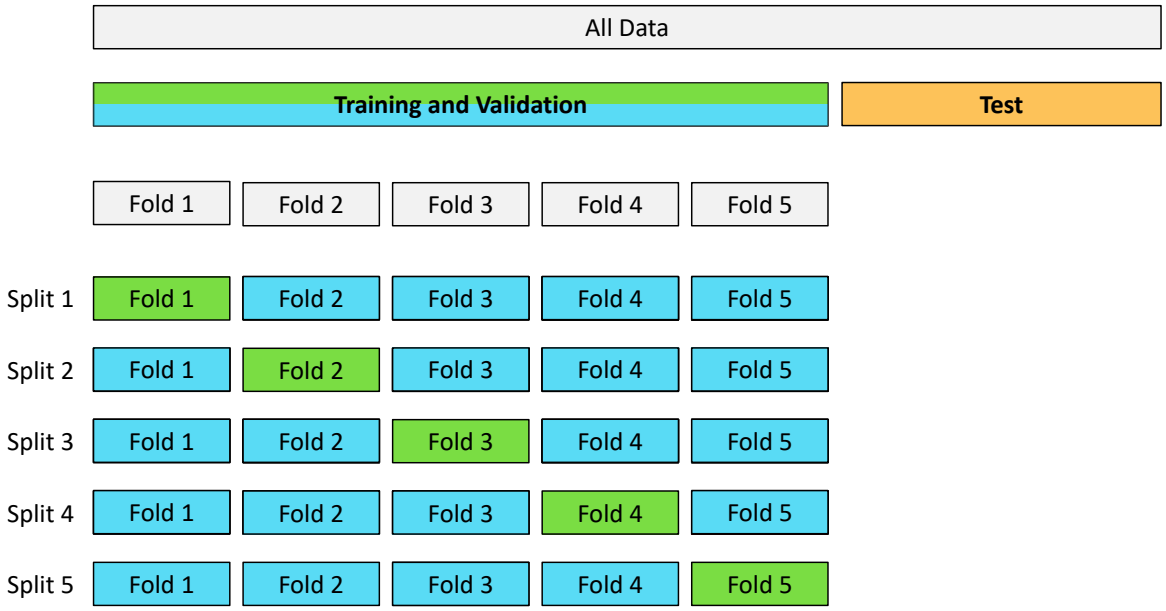


Figure 1.1: Overview of CV scheme. The training set is split into k smaller sets for training and validation. A model is trained using $k-1$ of the folds as training data; the resulting model is validated on the remaining part of the data. The procedure is followed for each of the k-folds. The performance measure reported by k-fold CV is then the average of the values computed in the loop.

where reserving data for independent test sets is not feasible. The method of nCV is relatively straightforward as it merely is a nesting of two k-fold CV loops: the inner loop is responsible for the model selection and the outer loop is responsible for estimating the generalization accuracy (Figure 1.2).

In the sphere of DL, the main power of a Deep Neural Network (DNN) lies in its deep architecture (Szegedy et al., 2015; Zeiler and Fergus, 2014), which allows for extracting a set of discriminating features at multiple levels of abstraction. However, training a DNN from scratch (also called full training) is not without complications. First of all is the scarcity of data in the medical context which leads to overfitting and convergence issues, whose resolution frequently requires repetitive adjustments in the architecture or learning parameters of the network to ensure that all layers are learning with comparable speed. To reduce the problem of overfitting due to small training samples, another promising AI technique has been developed in the literature called Transfer Learning (TL) which allows transferring knowledge of a pre-trained Neural Network (NN) from a source task into a target task (Durgut et al., 2022). In this regard, a model is adapted reusing the pre-trained NN weights, obtained from the source domain rich in samples, rather than starting the training from scratch within the small target domain. Often to improve the output it is convenient to make a fine-tune of the weights on the target task (Figure 1.3). There

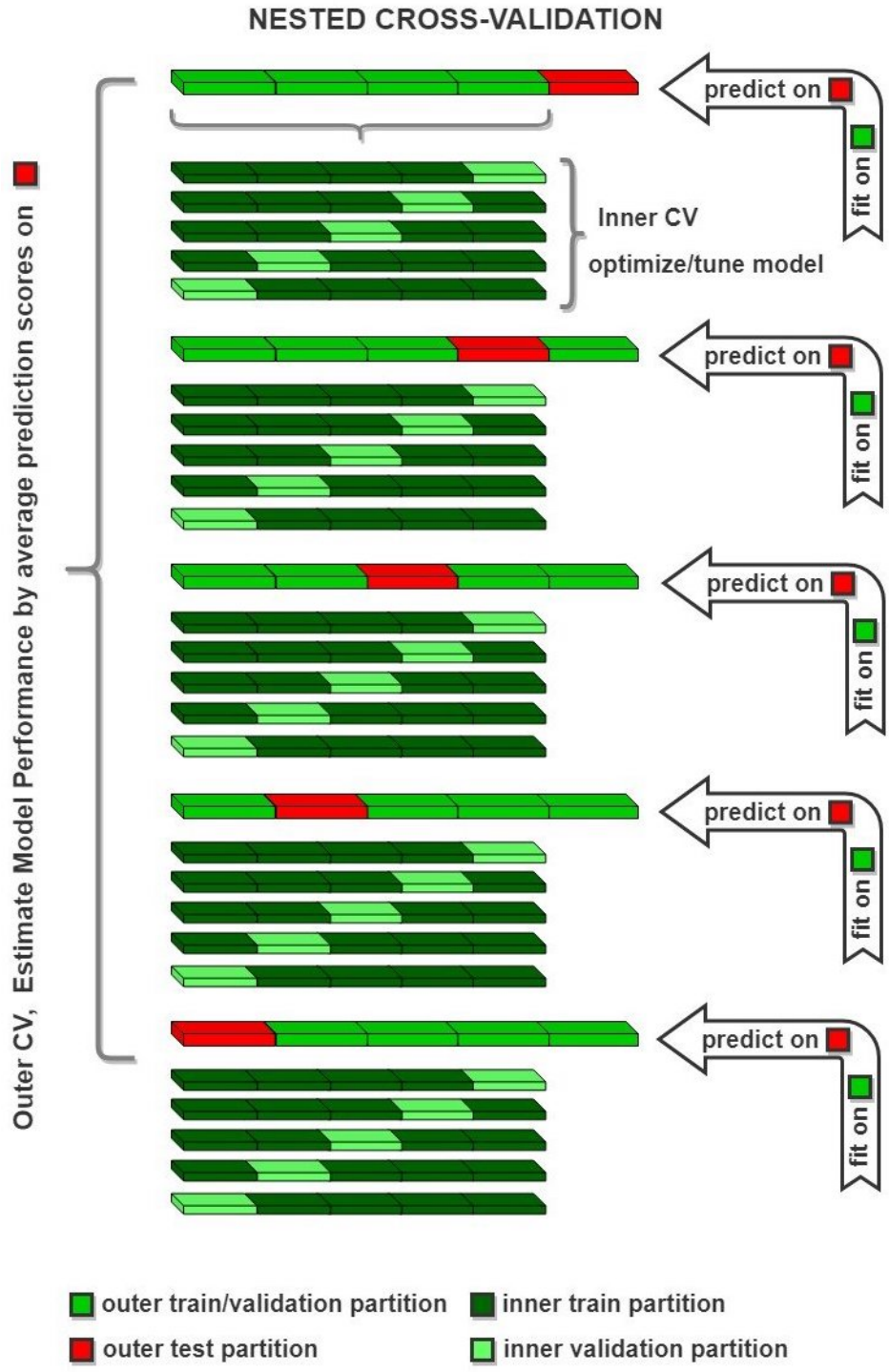


Figure 1.2: Overview of nCV scheme. Model optimization is performed via inner CV on the Train/Validation partition of each outer fold, where the optimally tuned model is selected based on average performance across the Validation inner partitions. The selected as optimal models are subsequently fit on the entire Train/Validation partition of the particular outer fold and deployed on the respective Test partition. Average Test scores and standard deviation across outer folds, provide estimates of model performance and generalization, respectively (Lavasa et al., 2021).

are various strategies, such as training the whole initialized network or "freezing" some of the pre-trained weights. With this in mind, the pre-trained network can be fine-tuned in a layer-wise manner, starting with tuning only the last layer (shallow tuning), then tuning all layers (deep tuning). In general, the early layers of a DNN learn low level image features, which are applicable to most vision tasks, but the late layers learn high-level features, which are specific to the application at hand. Therefore, fine-tuning the last few layers is usually sufficient for transfer learning. However, if the distance between the source and target applications is significant, one may need to fine-tune the early layers as well. Therefore, an effective fine-tuning technique is to start from the last layer and then incrementally include more layers in the update process until the desired performance is reached.

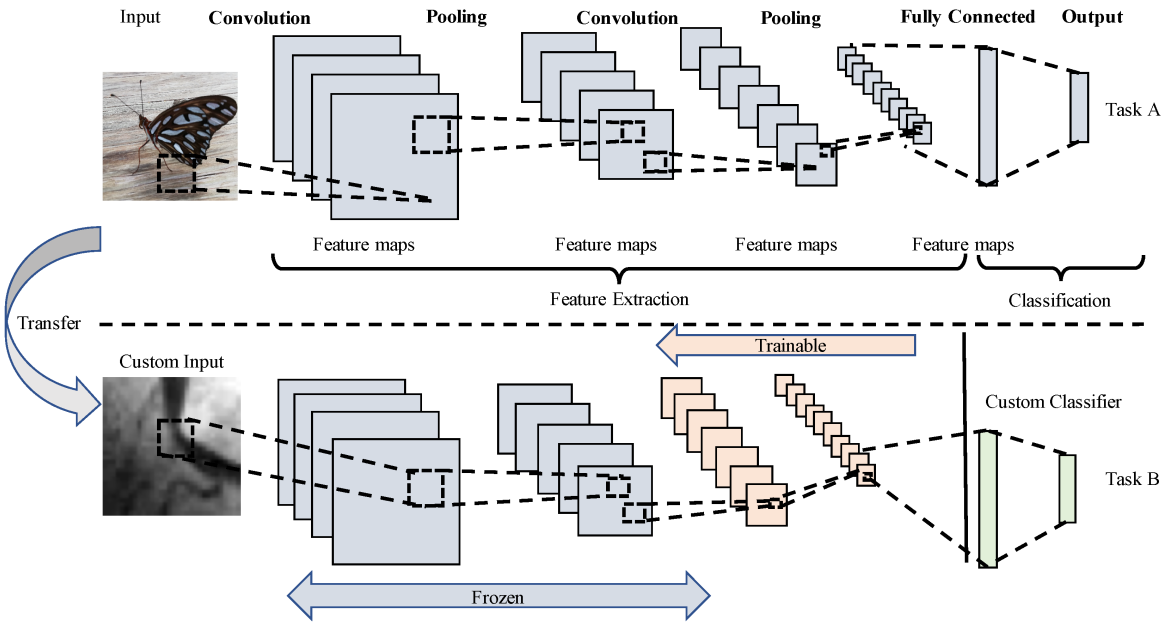


Figure 1.3: Transfer learning scheme. The last layers are replaced with custom layers for the proposed problem. The fine-tuning is carried out in the top-level layers, while the rest of the transfer networks remains temporarily idle or frozen (i.e., the weights stay unchanged during the optimization process) (Ovalle-Magallanes et al., 2020).

Data augmentation is an alternative method to training with more data, which involves the generation of more samples by applying simple or more sophisticated expedients. They can be cheap transformations (Figure 1.4) like affine transformations, cropping, adding noise, zooming, histogram equalization, sharpening, adjusting contrast or more expensive transformations such as deformable transformations (Chlap et al., 2021; Nalepa et al., 2019). While an innovative and very expensive resource regards the use of the Generative Adversarial Network (GAN) (Yang et al.,

2022; Han et al., 2019). The GAN model architecture consists of two deep learning networks (Figure 1.5), namely a generator that captures data distribution and a discriminator that tries to categorize the incoming input as a real or fake example (Goodfellow et al., 2014a). The learning procedure involves an adversarial process where the generator and the discriminator compete for one against the other in the creation of new synthetic images.

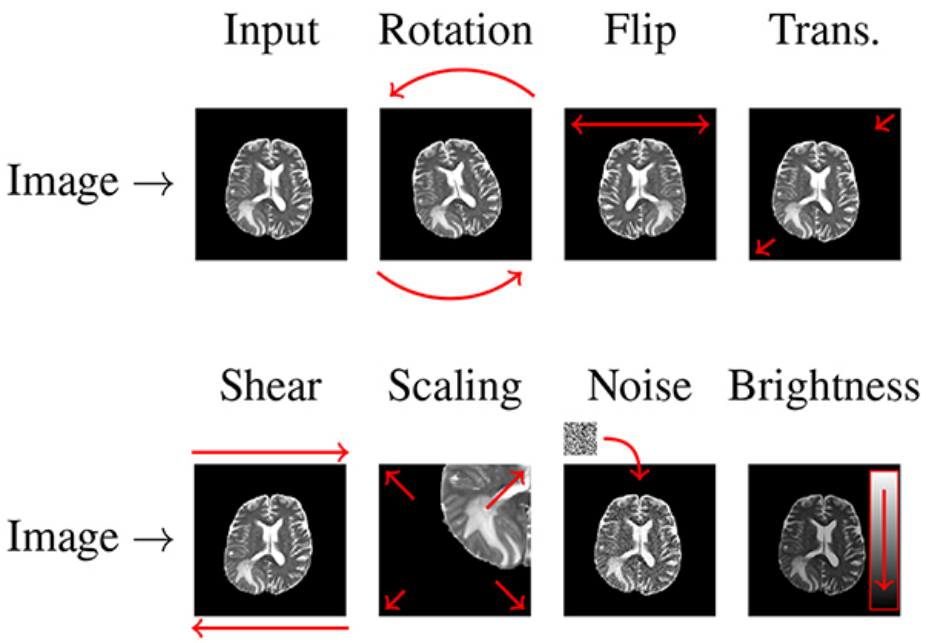


Figure 1.4: Applying affine and pixel-level transformations can help significantly increase the size of training sets. In this example, new images based on the original MRI have been generated (adapted from (Nalepa et al., 2019)).

In addition to what has already been announced, there are other challenges and pitfalls that undermine the success of a training and the achievement of a reproducible and generalizable AI model. Adequate curation, analysis, labeling, class imbalance, data leakage, external independent test, ethical issues and costs are detrimental for AI performance and critical to achieving high-impact clinically meaningful AI algorithms, if not properly accounted for (Lemaître et al., 2017; Buda et al., 2018; Dawud et al., 2019). Reliability and reproducibility of the results are mandatory for medical applications based on AI. Training AI models with limited annotated data samples poses specific challenges on the robustness and generalization ability of AI models. Specific guidelines should be defined regarding the definition of efficient training algorithms and rigorous cross-validation protocols either to enable the use of AI techniques in case of limited data availability for a specific study, or to discard

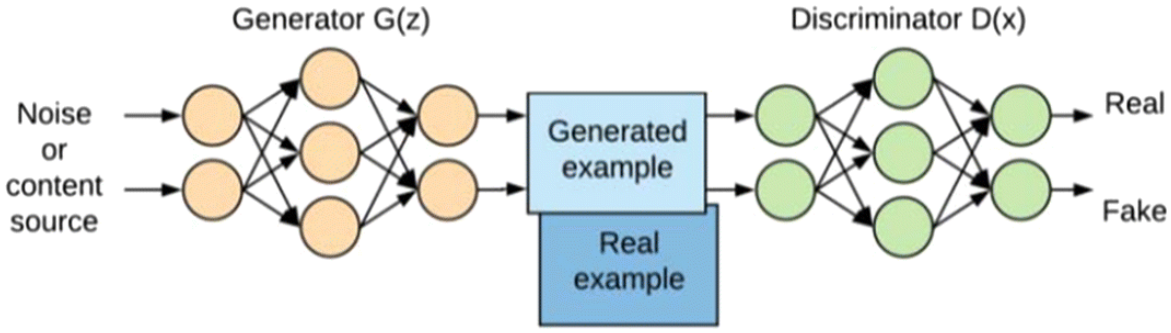


Figure 1.5: Illustration of GAN concept. The overall idea is to use two adversarial networks ($G(z)$ and $D(x)$), where one generates a photorealistic image in order to fool the other net (generator $G(z)$) trained to better distinguish fake images from the real ones (discriminator $D(z)$). In other words, the generator task is to minimize a cost function $V(D, G)$ (for example maximum likelihood), while discriminator needs to maximize it (Mikołajczyk and Grochowski, 2018)

the possibility of using them.

In the described framework, it falls the present thesis. It addresses the possibility, with proper precautions, of using artificial intelligence algorithms within small and real hospital database to fulfill specific clinical needs and to support clinical decision-making process. The thesis is organized into six parts.

- Chapter 2 explains and comments the fundamental steps which I followed for preparing medical images datasets to AI algorithm development. They can be summarized as follows: ethical approval, data access, querying data, data de-identification, data transfer, quality control, structured data and label data.

- Chapter 3 offers an overview of what was done in literature with small medical images databases. In particular, a systematic review was conducted to provide a comprehensive survey of recent advances in this field.

- Chapter 4 is related to the development of a radiomic-dosiomic workflow combined with machine learning algorithms in order to clustering patients in an unsupervised fashion, with special focus on exploiting Magnetic Resonance (MR) and dose images, as biomarkers of radio-induced neurotoxicity in pediatric patients affected by medulloblastoma.

- Chapter 5 and 6 present the use of artificial intelligence for the identification, segmentation and quantification of $COVID-19$ pulmonary lesions. The limited data availability and the annotation quality are relevant factors in training $U-net$ based algorithms. The effects of using multiple public datasets of $COVID-19$ CT scans, heterogeneously populated and annotated according to different criteria were investigated.

- Chapter 7 refers to a still-ongoing project that attempts to describe the prediction of patients' outcome with soft tissue sarcomas to radiotherapy, in terms of the development of distant metastases. Transfer learning and fine-tuning have been investigated to perform domain adaptation on a smaller private dataset.
- Finally, in Chapter8, conclusions and future works are presented.

This work has been carried out within the AIM and next AIM projects funded by INFN (CSN5, $2019-2021$ and $2022-2024$, respectively). The AIM and next AIM projects aim to exploit the expertise in advanced data processing and analysis techniques handled by National Institute of Nuclear Physics (INFN) and associated researchers on medical data processing and enhancement, and turn it in the development of advanced and effective analysis instruments to be eventually clinically validated and translated into innovative products for precision medicine. The AIM projects aspire to focus efforts on (i) the exchange of know-how on algorithms, data access, and codification of common problems, (ii) the training of young researchers on advanced cross-disciplinary problems, and (iii) the dissemination of acquired know-how to the scientific communities to achieve specific results on clinically-driven challenges. One of the objectives of the AIM project is to implement, optimize, develop and test ML, DL and network-based algorithms. In particular, to put in place predictive medicine solutions and it includes the prediction of treatment outcome in oncologic patients. While, for the next AIM project one of the purposes is to address the following specific challenge related to methodological aspects of the application of AI in medicine: how to manage limited datasets with AI techniques (no-so-big dataset).

Chapter 2

Medical Databases

The leading hurdles to development and implementation of AI algorithms in the clinical setting include availability of sufficiently large, curated, and representative training data that includes expert labeling (i.e. annotations). To make matters worse, medical imaging datasets acquired in clinical practice are often incomplete, and this could limit the applicability of models that instead require undamaged multiple modalities as input. These obstacles lead to possible biases of which one must be aware, which may affect generalizability of AI algorithms. Moreover, supervised AI methods for evaluation of medical image require a curation process for data to optimally train, validate, and test algorithms. Currently, most research groups have limited data access based on small sample sizes and/or from small geographic areas (Brehaut et al., 2006). In addition, the preparation of data is a costly and time-intensive process, the results of which are algorithms with limited utility and poor generalization.

Nowadays to alleviate the problem of data access in medical research and in large amounts, an increasing number of data sets has been open sourced. Data sets are available in a wide range of domains from brain MRI (Bakas et al., 2017; Di Martino et al., 2014; Jack et al., 2008; Mennes et al., 2013; Menze et al., 2015; Van Essen et al., 2013), breast imaging (Lee et al., 2017; Xi et al., 2018), chest radiographs (Irvin et al., 2019; Wang et al., 2017b), and others. Unfortunately, these databases concern common and routine pathologies, in fact they manage to reach large numbers also and above all for this reason. Often clinicians would like to be able to stratify or isolate rare diseases and therefore poorly known and uncommon. There are other important limitations as well. First, there is a wide variety of number and quality of images and availability of metadata and clinical information. Second, some open-source data sets contain low-quality images, lack expert labeling or data curation (Langlotz et al., 2019). Therefore, most academically developed AI algorithms in medical imaging,

including mine, have been trained, validated, and tested with local data from a single institution.

In this section, the fundamental steps which I followed for preparing medical images data sets to AI algorithm development will be described. One more footnote should be added. The creation of a medical database involves many professional roles: administrators, technicians, medical physicists, physicians and researchers. From this point of view, my task was also to interface with all these experts and act as a bridge between them.

2.1 Data Preparation Overview

Before medical images can be used for the development of an AI algorithm, certain steps need to be taken. Typically, approval from the local ethical committee is required before medical data may be used for development of a research. In my case, however, retrospective studies have been conducted and therefore existing data are used. Because the patients in this type of study do not need to undergo any additional procedures, explicit informed consent is not formally required.

After ethical approval, relevant data needs to be accessed, queried, properly de-identified, and securely stored. Any protected health information needs to be removed both from the Digital Imaging and Communications in Medicine (DICOM) metadata, as well as from the images (`https://mircwiki.rsna.org/index.php?title=MIRC_CTP`). The quality and amount of the images vary with the target task and domain. The next step is to structure the data in homogenized and machine-readable formats (Harvey and Glocker, 2019). The last step is to link the images to ground-truth information, which can be one or more labels or segmentations. The entire process to prepare medical images for AI development is summarized in Figure 2.1.

2.2 Accessing and Querying Data

Fortunately, for the role I hold within the hospital and having received approval from the ethics committee, I have direct access to medical imaging data through the Picture Archiving and Communication System (PACS). In fact, only accredited professionals can access to PACS environments. Once data are accessible, different strategies are available to search for medical images and clinical data. The most efficient and immediate search queries take place through the use of names and identification codes of patients. It is necessary to underline the fact that the data to

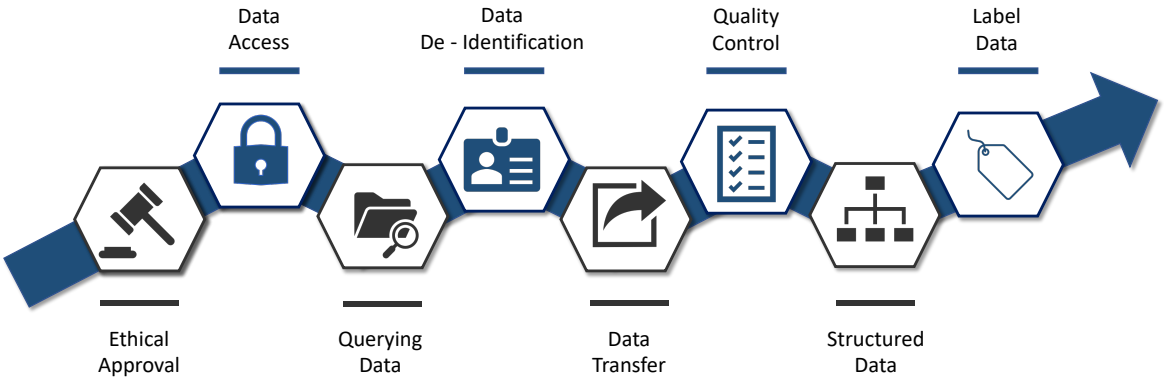


Figure 2.1: Diagram shows process of medical image data handling.

search and extract necessary to develop the project often do not belong to the same PACS storage environment, but rather to two or three different repositories and the only common denominator that allows a transversal search is the patient's name. This happens because different departments can, and indeed often have, different storage systems that do not communicate with each other.

2.3 De-Identification

Although written informed consent from patients is not always necessary, retrospectively gathered data require proper de-identification. Sensitive information includes but is not limited to name, medical record number, and date of birth. Identifiable information is commonly present in the DICOM metadata (header) and multiple tools are available to automatically remove this information (Aryanto et al., 2015); I relied on MATLAB (MATLAB 2020*b*, The MathWorks, Inc., Natick, Massachusetts, United States). For spreadsheets I have adopted anonymization with *k*-anonymity, which transforms an original data set containing protected health information to prevent potential intruders from determining the patient's identity (El Emam and Dankar, 2008). Other strategies that can be employed act in such a way that the DICOM metadata is often removed completely or converted to another format such as NIfTI (Neuroimaging Informatics Technology Initiative) which retains only voxel size and patient position. Totally removing the DICOM metadata certainly prevents privacy issues but reduces the value of data, because metadata contain important information for AI algorithm development.

2.4 Data Storage

In my situation, data are transferred to a local data storage and more precisely on two different hard drives to ensure their use and backup. The advantages of this type of archiving include data security and availability but as you can well imagine in this way the possibility of real-time sharing with colleagues is lost.

2.5 Quality control and Structured data

It is fundamental in a database to have quality data; it is not enough to have a large amount of information available if it is of little value. Therefore, it is necessary to carry out a quality control of the collected images. The first check to be implemented concerns the integrity of the files. Often the data is stored in archives and when they are restored, they can become corrupted. Even if at first glance they seem flawless, reading them returns an error which renders them unusable. Or the DICOM header may be corrupted and essential information for the correct use of the images are lost. The second step concerns the quality of the image itself, such as an excessive presence of noise or artifacts due to movement, breathing, presence of metal prostheses or contrast medium. Subsequently, a standardization of the nomenclatures within the database must be obtained, eliminating variations in terminology and building the mapping to a common controlled terminology. This is recurrent in the segmentation phase in radiotherapy where the process is performed by different physicians. Even if they have a common vocabulary it is usual to find differences in the identification of the same Region Of Interest (ROI), even simply deviations between uppercase and lowercase letters.

Making the information stored in the database easily accessible to the AI model developer, the data must be arranged in such a way to create a set of structured data, i.e. organized in an orderly manner, according to a set of predetermined rules. For example, if researchers are dealing with multimodality images from different scanners (i.e. CT and MRI), for each patient they should sort the images with the same orientation and direction (head first or feet first protocol, supine or prone); or if the individual slices are taken into consideration, it is useful to associate each one with its own segmentation or binary mask if available. Nor is it essential to hook the appropriate label to undertake supervised learning. Thus, objects structured in an equal and repetitive way are obtained, which can be easily implemented in a training, validation and test algorithm.

2.6 Appropriate Label

Current AI algorithms for medical image classification tasks are generally based on a supervised learning approach. This means that before an AI algorithm can be trained and tested, the ground truth needs to be defined and linked to the image. The term ground truth typically refers to information acquired from direct observation (such as biopsy, autopsy or laboratory results). For this reason, in my studies, the term 'image labels' will refer to a (retrospective) annotations performed by medical experts such as radiologists or radiotherapists. These annotations have to be understood as the result of a condensation of different diagnoses and information, such as free-text radiology report or expert consensus and interpretation (Hwang et al., 2019). Even simply because, in the clinical setting, the final annotation may require further confirmations in addition to the diagnostic report such as a pathologic or surgical report, clinical outcome, follow-up (i.e., cancer type, metastases development, induced toxicity occurrence, etc.).

As far as I'm concerned, since I don't have the necessary skills or knowledge to be able to interpret the clinical reports and translate them into a format suitable for the development of a medical imaging classification AI model, I obtained the appropriate labels for the patients included in the databases by interviewing the referring physician supervising the study. As to be expected, manual labeling required a substantial effort.

Chapter 3

State-of-the-art

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

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

[1] National Institute for Nuclear Physics (INFN), Florence Division, Florence, Italy
[2] University of Florence, Florence, Italy
[3] Department of Soil, Plant and Food Sciences, University of Bari Aldo Moro, Bari, Italy
[4] National Institute for Nuclear Physics (INFN), Bari Division, Bari, Italy
[5] National Institute for Nuclear Physics (INFN), Pisa Division, Pisa, Italy

I contributed to think and design the review. I equally collaborated to screening the abstracts of potentially eligible studies and to the complete reading of the admitted articles. I significantly contributed (80%) to extract the data from the study reports. I also drafted the review (100%).

Abstract

Background: 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.

Methods: 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 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 statements (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis).

Results: To solve the small data issue transfer learning technique, basic data augmentation and GAN 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 4 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).

Conclusions: The goal of this review is to provide a comprehensive survey of recent advancements in dealing with small medical images sample size. Transparency and improve quality in publications as well as follow existing reporting standards are also supported.

Keywords: *Small Data, Artificial Intelligence, Medical Imaging, Transfer Learning, Data Augmentation, Classification*

3.1 Introduction

Data-driven intelligent models have gained immense popularity in recent years, achieving amazing performance in various fields of our daily life. 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 (Romero et al., 2020; Swati et al., 2019). 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 (D'souza et al., 2020; Ubaldi et al., 2021).

In numerous real-world applications, the number of samples in a dataset can be relatively limited, constrained by the complexity, ethnicity, high cost or they can be difficult to obtain in practice, leading to sharply decreases in 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 (D'souza et al., 2020). 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 (Vabalas et al., 2019). 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 (Xu et al., 2021; Ubaldi et al., 2021). 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 (Trivizakis et al., 2019; Ayana et al., 2022).

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 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 (Han et al., 2021). 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 provide a potential data source to train models specific to medical applications.

The small data issue can be faced mainly 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 by mimicking the distribution of the original dataset, and 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 models by reducing the possibility of overfitting (Adedigba et al., 2022; Gatidis et al., 2015).

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 the generalization.

One such more objective and promising technology that has recently been introduced for data augmentation, is the Generative Adversarial Network (GAN) which involves generative models and adversarial learning (Goodfellow et al., 2014b; Pan et al., 2022). 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 of the original training data (Levine et al., 2020; Gheshlaghi et al., 2021; Shi et al., 2020; Zebin and Rezvy, 2020).

Transfer learning

Another possible way to face the small sample size problem is the transfer learning, that is to use a pre-trained network which cleverly applies the 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 datasets and then

uses the trained model either as an initialization or as it is for a new task (Alruwaili and Gouda, 2022; Horry et al., 2020; Bahgat et al., 2021).

The idea is to initialize the neural network with the weights trained from some previous task and then fine-tune the parameters within the current task when the current task has insufficient training data. This approach provides a reasonable initial state and may speed up the training of the model, 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 a pre-trained network: (1) reusing the parameters of a pre-trained deep neural network directly to initialize the new network and fixing without retraining, called freezing. (2) reusing the parameters of the pre-trained deep neural network directly to initialize the new network and then 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 (Romero et al., 2020).

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 (Russakovsky et al., 2015), 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 contexts which have a small number of image data and they 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 can 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 (Aderghal et al., 2020; Sha et al., 2019; Sanchez et al., 2022).

The aim of this work 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. 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.

3.2 Methods

PRISMA

This manuscript has been prepared according to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines and a checklist is available in the supplementary material (Moher et al., 2009).

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 July 31, 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). 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, database description and data preprocessing.

Screening of collected studies

After removal of clearly irrelevant records, two reviewers independently screened the abstracts of 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, to 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.

Adherence to reporting standards - TRIPOD

We evaluated the quality of the studies according to Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement (Collins et al., 2015). This statement rates the transparency of the reporting of a prediction model study regardless of the survey methods used and in all medical settings (Moons et al., 2015). It is composed of a 22 items checklist (37 total points when all sub-items 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 the studies broadly conformed to the reporting recommendations included in TRIPOD, and not the detailed granularity required for a full assessment of adherence (Heus et al., 2019).

Data synthesis

Aware of heterogeneity of specialties, metrics and outcomes, we reported in Table 3.1 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 of code and database.

3.3 Results

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 3.1).

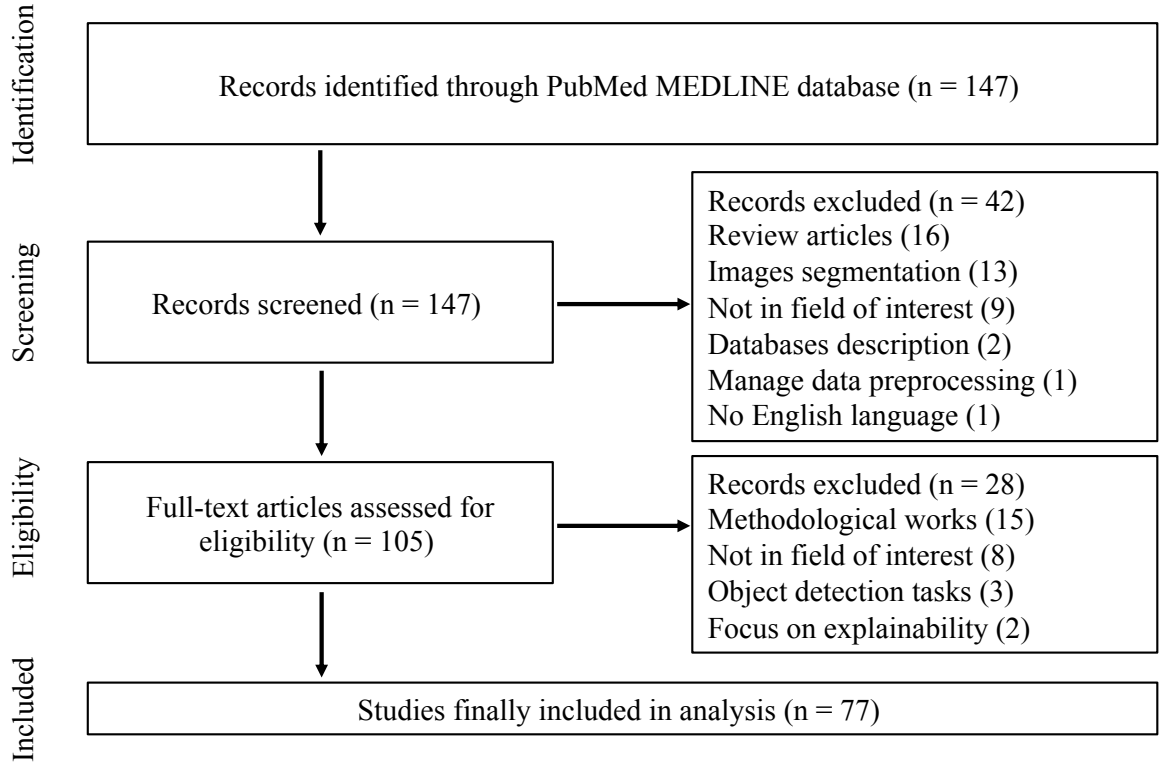


Figure 3.1: Flow-chart of article selection based on PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines.

General characteristics

Table 3.1 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 models and only 2 (3%) with prognostic models. Most of the works make use of deep learning techniques (86%), only 6% apply only traditional 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 3.2*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), Head & Neck (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 16600 ± 45700 samples (mean $\pm$ one standard deviation), a minimum of 16 and a maximum of 299000 (Figure 3.2*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 3.2*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 3.2*d*). Concerning reproducibility, data are public and available in 47 studies. In 25 analyses the collected data are private and 7 operate over both types of databases. Fifty per cent 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 (3.1).

To solve the small data issue transfer learning technique, basic data augmentation and GAN were 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 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.

Fiftythree 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 19 others did not report on funding.

Table 3.1: 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 (%)*	Test Accuracy	Test AUC	# of classes	Imaging modality	Anatomical region	Database type	# of used database	External Test	Sharing code
Swati	2019	X			233		5-fold		0.95		3	MRI	Brain	Public	1		X
Abbasi	2021	X	X		26431	31231	5-fold		0.82	0.89	2	Retinal images	Eyes	Public	2		
Khan	2020		X		253			73:19:8	1.00	1.00	2	MRI	Brain	Public	1		
Romero	2020	X	X		112120			70:10:20		0.95	2	XRay	Lung	Public	4		
Horry	2020	X	X		2368	28560		80:20			2	Ultrasound, XRay, CT	Breast	Public	4		
Alzubaidi	2021	X	X		143243	343243		80:20	0.98		4	Histologic	Skin	Public	2		X
Wodzinski	2020	X	X		174		5-fold		0.74		2	MRI	Brain	Private	1		
Hertel	2021	X	X		31595	37914		90:10	0.94	0.98	3	XRay	Lung	Public	5		
Baydilli	2020				263		10-fold	75:10:15	0.97	0.93	5	Histologic	White Blood Cells	Public	1		
Li	2019	X	X		15573	140157	10-fold	85:5:10	0.97	1.00	4	Retinal images	Eyes	Mix	3		
Xia	2018	X			299096			90:10	0.84	0.92	2	Histologic	Lymph node metastases	Public	2		
Shen	2020	X	X		668			73:10:17	0.96		3	XRay	Adenoid	Private	1		
Feng	2018				58000		10-fold		0.98		2	Histologic	Breast	Public	1		
Liu	2020	X		X	24000	27000		75:13:12	0.88	0.93	2	Histologic	Brain	Mix	2		X
Levine	2020	X	X	X	1022		10-fold		0.92	0.99	10	Histologic	Different tumors	Public	2		X
Ahn	2020	X	X		13986		5-fold	65:35	0.83	0.83	2	Histologic	Skin	Public	3		
Gheshlaghi	2021	X	X	X	13338	14838		70:30	0.90		2	Histologic	Breast	Public	1		
Montoya	2018				105			33:8:59	0.99		3	CT	Brain	Private	1		
Xia	2020				373			65:35		0.90	2	CT	Lung	Private	2		X
Liang	2021	X	X		100	640		80:20	0.91	0.91	2	MRI	Brain	Public	1		
Huynh	2016	X			607		5-fold			0.86	2	XRay	Breast	Private	1		
Zhang	2021				1870				0.57	0.60	6	MRI	Brain	Public	2		
Hu	2020	X			499		5-fold	55:45	0.71	0.75	2	MRI	Brain	Public	2		
Dai	2021	X	X		1714			80:10:10	0.95	0.98	3	MRI	Knee	Private	1		

* Respect to the original available data.

Table 3.2: 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 (%)*	Test Accuracy	Test AUC	# of classes	Imaging modality	Anatomical region	Database type	# of used database	External Test	Sharing code
Apostolopoulos	2020	X	X		216	2216	10-fold		0.75	0.81	2	SPECT	Heart	Private	1		
Bien	2018	X	X		2287			74:13:13	0.91	0.99	3	MRI	Knee	Mix	2	X	X
Fu	2020	X	X		1091			60:20:20	0.95	0.98	3	Histologic	Skeletal	Public	1		
Chougrad	2018	X	X		6116		5-fold		0.98	0.99	2	XRay	Breast	Public	3		
Hu	2021	X	X		217		5-fold		0.92	0.99	2	MRI	Prostate	Private	1		
Shi	2020	X	X	X	1937		10-fold	84:9:7	0.92	0.95	2	US	Thyroid	Private	1		
Ye	2020	X	X		650		10-fold		0.92	0.96	2	CT	H&N	Private	1		
Zhou	2021	X	X		616			75:25	0.83		2	CT	Liver	Private	1		
Yi	2019	X	X		3034			70:10:20		1.00	2	XRay	Breast	Public	1		
Mutasa	2018		X		10289		8-fold	86:10:4				XRay	Skeletal	Mix	2		
Mzoughi	2020		X		284			75:25	0.96		2	MRI	Brain	Public	1		
Wang	2017	X	X		233		10-fold				5	CT	Lung	Public	1		
Samala	2020	X	X		3411	27288	4-fold			0.83	2	XRay	Breast	Mix	2		
Yi	2019	X	X		250	5760		64:12:24		1.00	5	XRay	Skeletal	Mix	2	X	
An	2021	X	X		954			80:20		0.91	5	Retinal images	Eyes	Private	1		
Owais	2020	X	X		52471		2-fold	50:50	0.96		37	Endoscopic	Gastrointestinal	Public	2		X
Samala	2021	X			4577		4-fold	70:10:20		0.72	2	XRay	Breast	Mix	2		
Cogan	2019	X	X		8000	27200		85:15	0.98	0.94	8	Endoscopic	Gastrointestinal	Public	1		
Apostolopoulos	2020	X	X		3905		10-fold		0.99		2	XRay	Lung	Public	3		
Choi	2017	X	X		279	10000	5-fold	70:30	0.87		2	Retinal images	Eyes	Public	1	X	
Zong	2020	X	X		528		10-fold	60:40	0.85	0.84	2	MRI	Prostate	Public	2		
Zebin	2021	X	X	X	802	902	5-fold	80:20	0.97		3	XRay	Lung	Public	2		X
Aderghal	2020	X	X		1551	184320		70:20:10	0.92	0.94	2	MRI	Brain	Public	1		
Uemura	2021	X		X	333	12407		80:20		0.89	2	CT	Colon	Private	1		
Oakden	2017				15957		6-fold		0.69	0.68	2	CT	Lung	Private	1		
Nabizadeh	2021				22232				1.00		2	XRay	Lung	Public	3		

* Respect to the original available data.

Table 3.3: 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 (%)*	Test Accuracy	Test AUC	# of classes	Imaging modality	Anatomical region	Database type	# of used database	External Test	Sharing code
Wang	2020	X	X	X	206	48900	3-fold		0.61	0.83	3	CT	Lung	Private	2		X
Haga	2018				40		30-fold	70:30	0.66	0.73	2	CT	Lung	Private	1		
Fantini	2021	X	X		10880	76800	3-fold	70:30	0.95		2	MRI	Brain	Private	4		
Trivizakis	2019		X		130	796		57:25:18	0.83	0.80	2	MRI	Liver	Private	1		
M Bahgat	2021	X	X		12933			85:15	0.99	1.00	4	XRay	Lung	Public	8		X
Zhang	2019		X		130		2-fold		0.83	0.86	2	CT	Colon	Private	2		
Toda	2021	X	X	X	66		3-fold		0.62		3	CT	Lung	Private	1		
Gatidis	2015				16		10-fold	94:6	0.89		2	MRI, PET	Prostate	Private	1		
Sha	2019				100			74:26	0.89	0.73	2	PET	Lung	Private	1		
Usman	2022	X			197087			80:10:10	0.89	0.87	14	XRay	Lung	Public	2		
Kaur	2022				2482			80:20	0.99	1.00	2	CT	Lung	Public	1		
Alruwaili	2022	X	X		99			70:10:20	0.90		2	XRay	Breast	Public	1		
Adedigba	2022	X	X		410	18200		80:15:5	1.00		6	XRay	Breast	Public	1		
Hashemzehi	2021				6328			80:20	0.93		3	MRI	Brain	Public	2		
Rocca	2021				30				0.93		2	CT	Liver	Private	1		
Suganyadevi	2022	X			7000		5-fold	70:10:20	0.99		2	XRay	Lung	Private	2		
Ahmad	2022	X	X	X	3064			60:20:20	0.96		3	MRI	Brain	Public	1		
Le	2022	X	X		669		5-fold	80:20	0.79	0.82	3	CT	H&N	Public	2	X	
Ayana	2022	X	X		21000	28920	nested 5-fold	70:15:15	0.99	1.00	2	Histologic, US	Breast	Public	3		
Cahan	2022	X	X		358			68:12:20	0.85	0.88	2	CT	Heart	Private	1		
Ho	2022	X	X		3783		10-fold	80:20	0.94		3	XRay	Lung	Public	5		
Muhammad	2022	X	X		19196	29576			1.00		2	XRay, CT	Lung	Public	3		
Sanchez	2022	X		X	2973		10-fold	94:6	0.98	0.96	2	XRay	Lung	Mix	2		
Sahoo	2022	X	X		85672			80:20	0.99	1.00	3	XRay, CT	Lung	Public	2		X
Zhang	2022	X	X	X	7254		5-fold		1.00	1.00	2	Retinal images	Eyes	Public	3		X
Ben Ahmed	2022	X	X		209	10000		80:20	0.74	0.70	2	MRI	Brain	Public	1		X
Ettehadi	2022				6720			60:20:20	0.98		4	MRI	Brain	Public	2		

* Respect to the original available data.

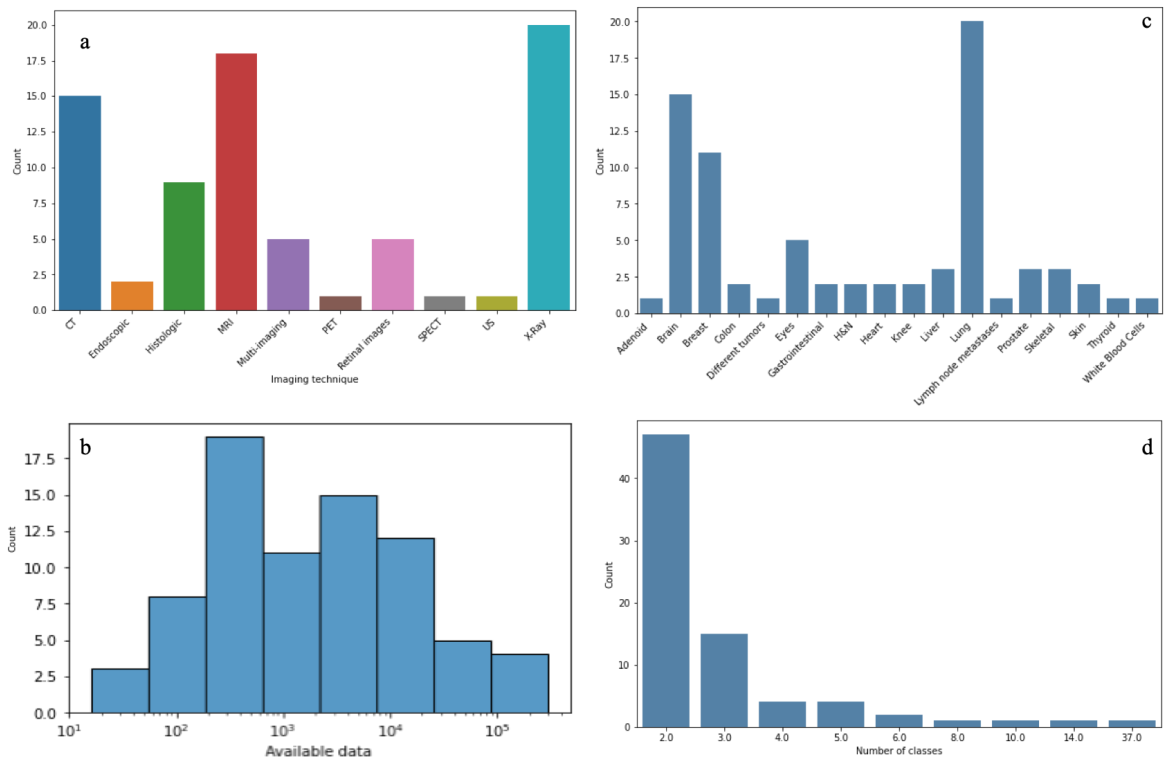


Figure 3.2: General characteristics concerning the imaging modalities (a), distribution of the number of samples in the databases (b), the most popular anatomical regions (c) and the preferred type of classification (d).

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 3.3). 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 (Figure 3.4). By comparing the performance metrics with respect to the use or not of this technique, differences can be noted (Figure 3.5). 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 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 ($p = 0.03$).

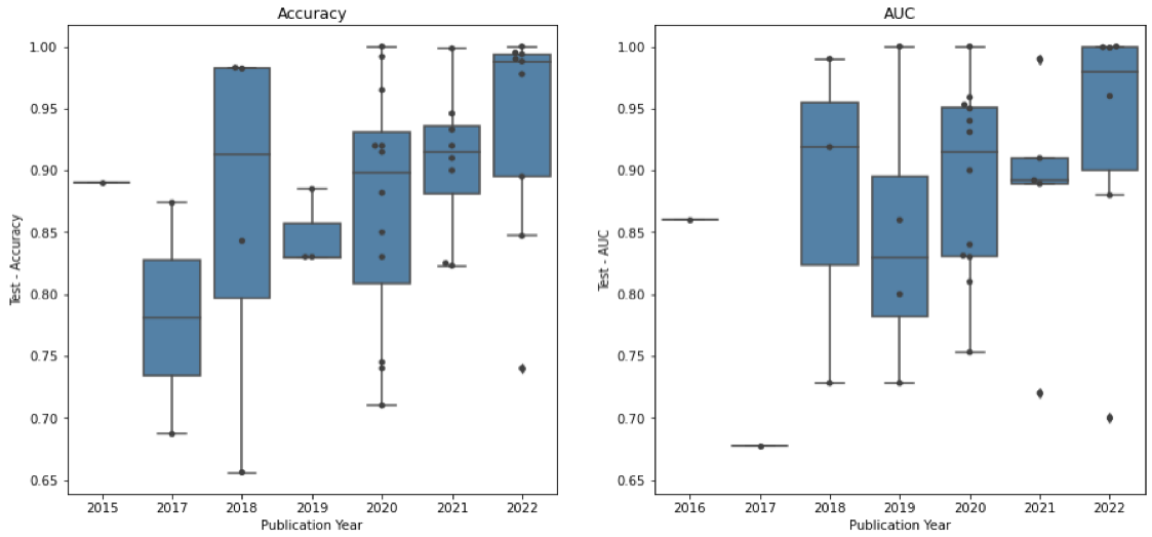


Figure 3.3: Performance of AI algorithms in terms of accuracy and AUC as function of publication year for binary classification studies

Adherence to reporting standards

Adherence to reporting standards less than 50% is present in 13 of 37 TRIPOD items (Figure 3.6). Overall, publications adhered to between 52% and 88% of the TRIPOD items: median 68%, interquartile range $61 - 71\%$, 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 comments: 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 been found because in the title the authors have 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 material.

For the moment, quantity and quality have not helped to improve performances (Figure 3.7). 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

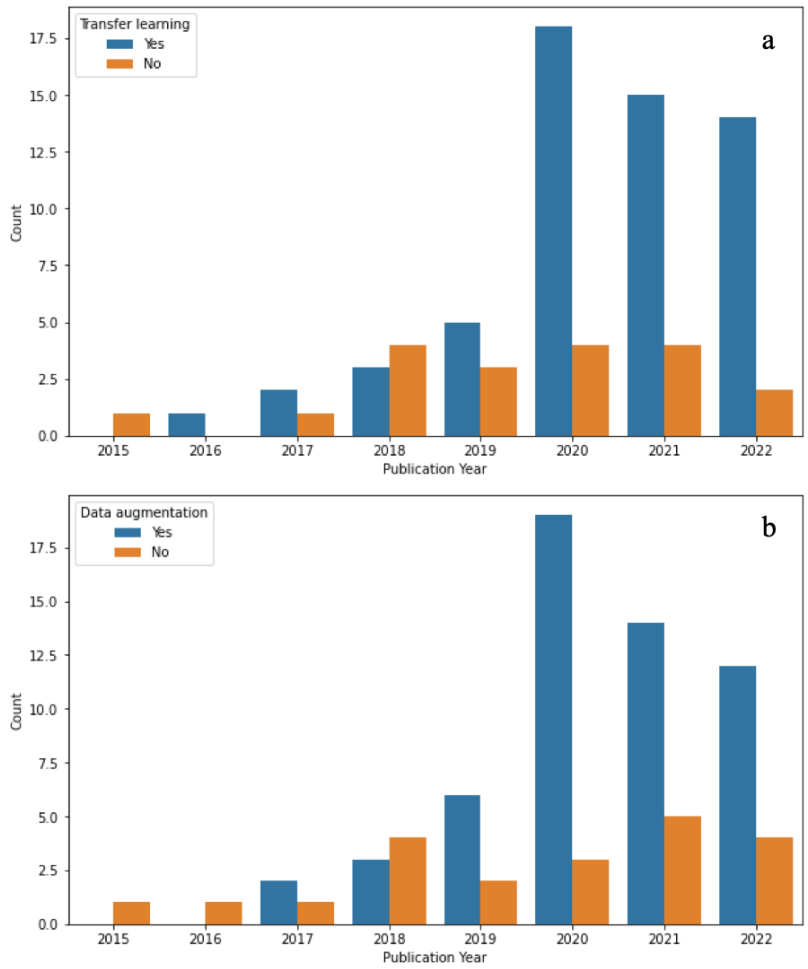


Figure 3.4: Use of transfer learning (a) and data augmentation (b) as function of publication year.

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 quality (TRIPOD index) by anatomical region, imaging technique and dataset origin can be found in the Supplementary Materials (3.2 and 3.3).

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

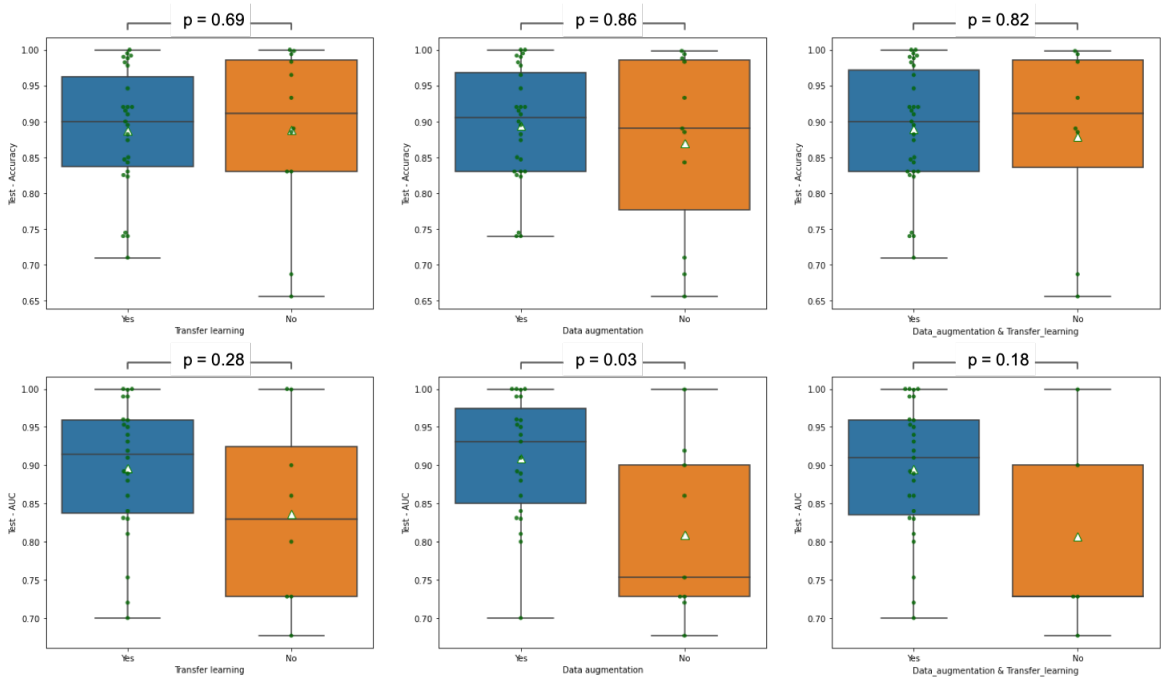


Figure 3.5: AI performances (accuracy top, AUC bottom) for binary classification studies; with and without transfer learning (first column), with and without data augmentation (second column), with and without both techniques (third column).

focus on improving design, validation, transparency and sharing (Le et al., 2022).

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 3.1, certain databases are difficult to classify as small in absolute terms having more than 100000 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 (DeVries and Taylor, 2017). 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 (Vabalas et al., 2019; Homeyer et al., 2022; Yu et al., 2022).

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

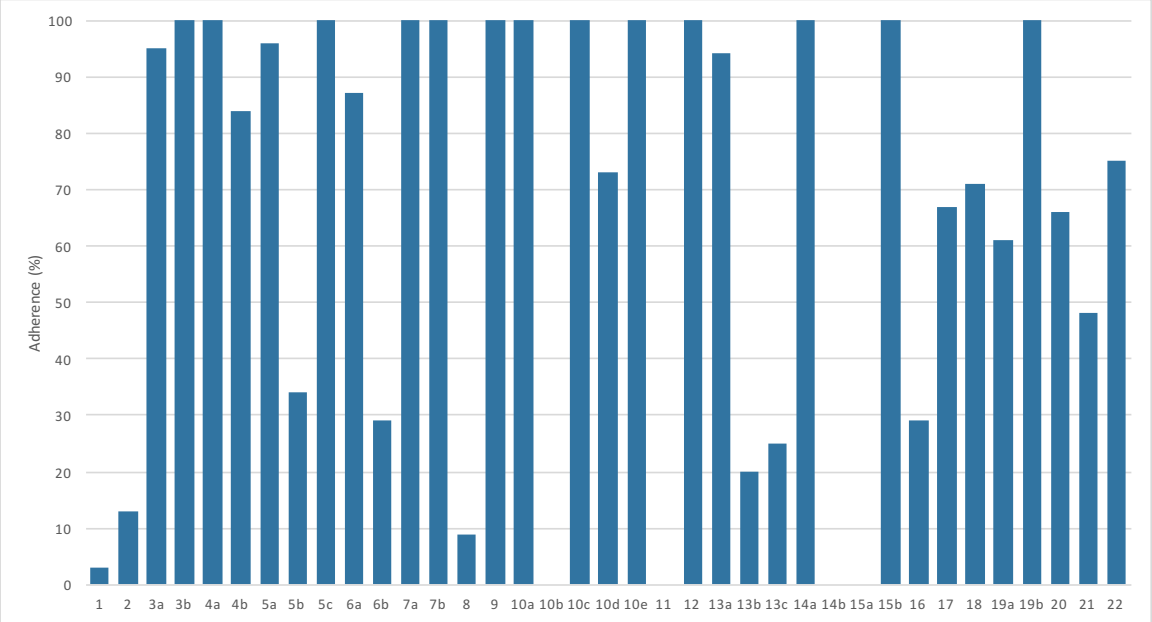


Figure 3.6: Completeness of reporting of individual TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) items.

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 (Heim et al., 2018). 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 intelligent research. To overcome the main drawbacks and pitfalls in this field,

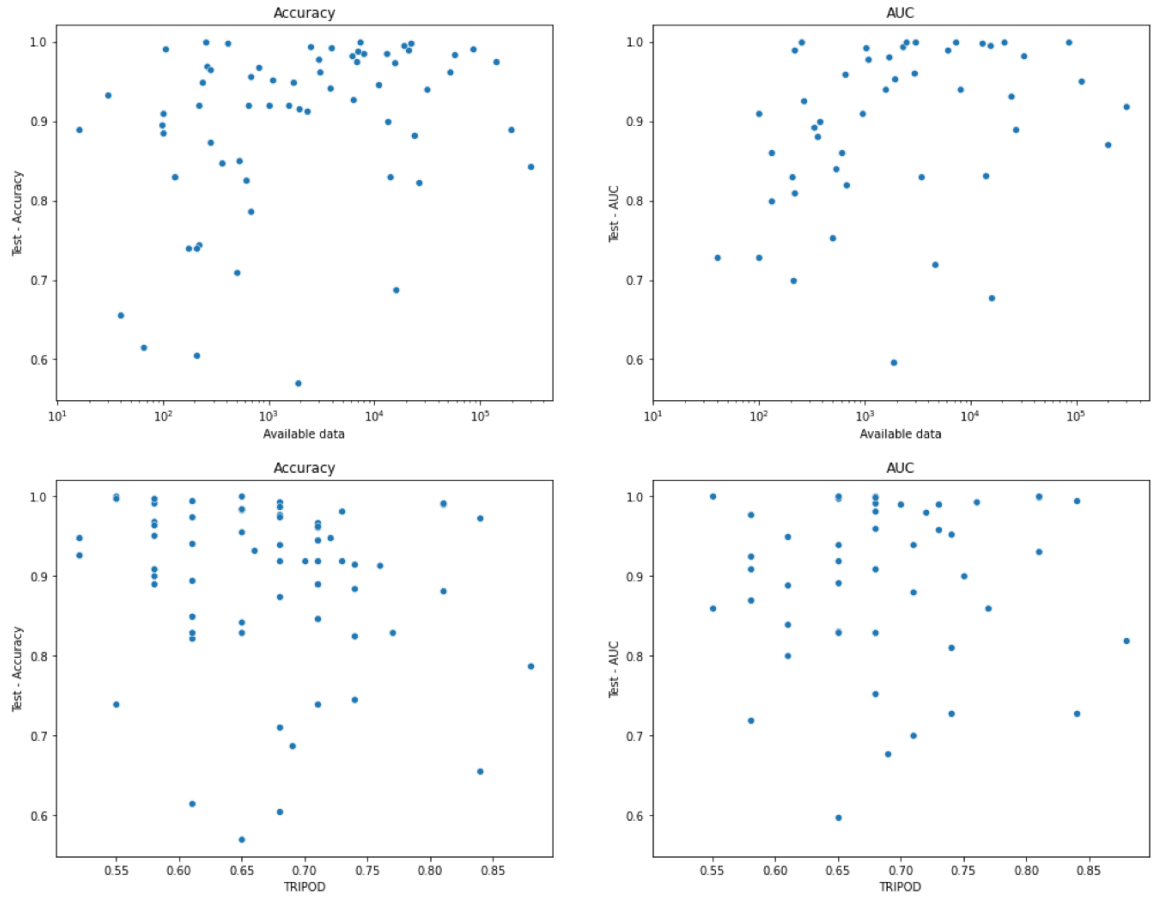


Figure 3.7: AI performances with respect to the quantity (Available data) and quality (TRIPOD) of the data.

reliable and efficient strategies must be considered and applied (Alzubaidi et al., 2021; Dai et al., 2021; Samala et al., 2020).

As the systematic review revealed, researchers rely 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. 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 (Alzubaidi et al.,

2021; Uemura et al., 2020; Usman et al., 2022). However, they will be validated, tailored, and improved to the specific application to achieve optimal results. 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.

An important 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. (2022) 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 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 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's 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. Isn't 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 or development process and helps

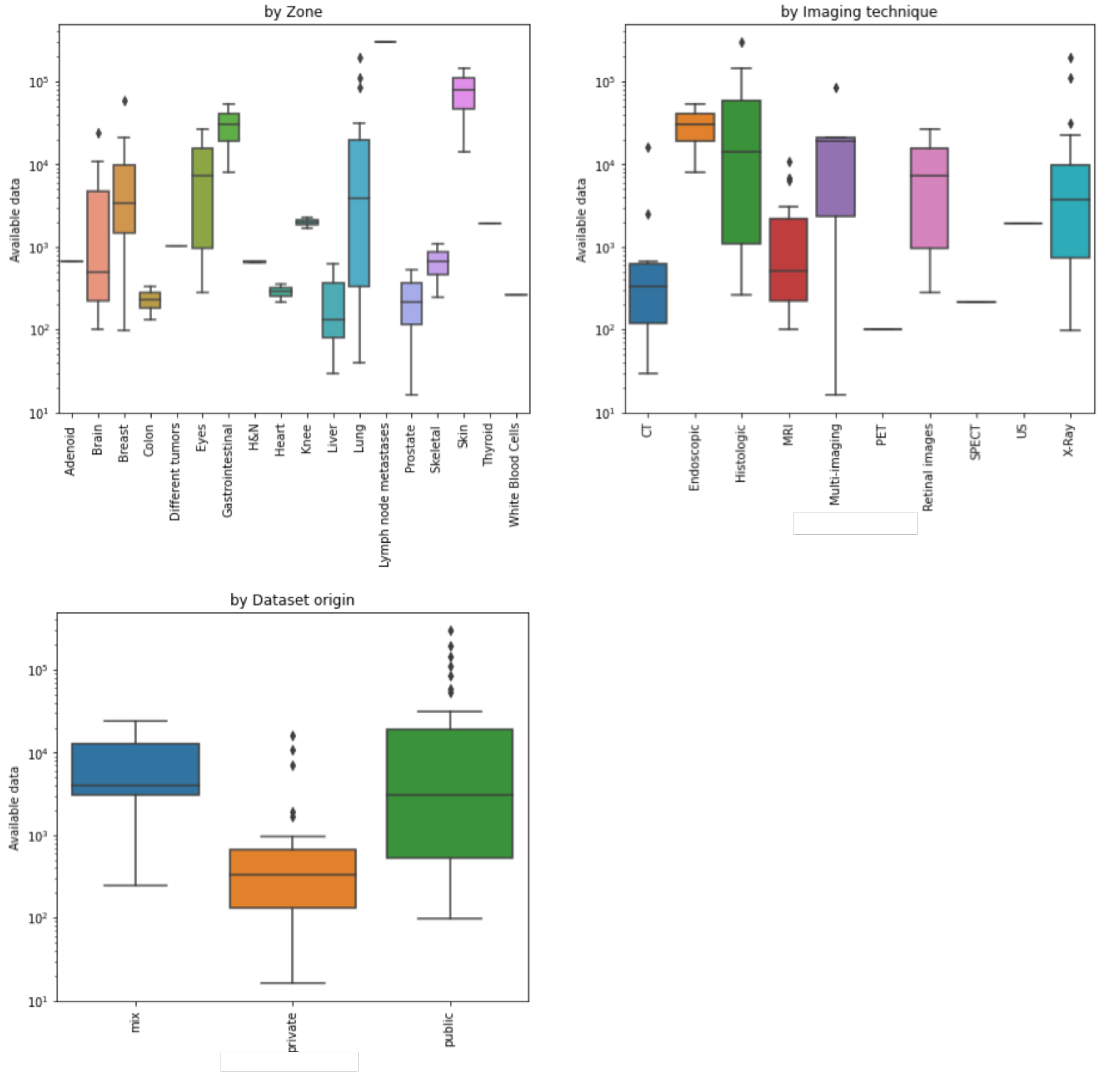
build trust among stakeholders, decision-makers, and the wider community. Overall, external validation is an important process for ensuring that models are performing as intended, and that the results are accurate and reliable against real-world data. In addition, it provides confidence in the decisions 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.

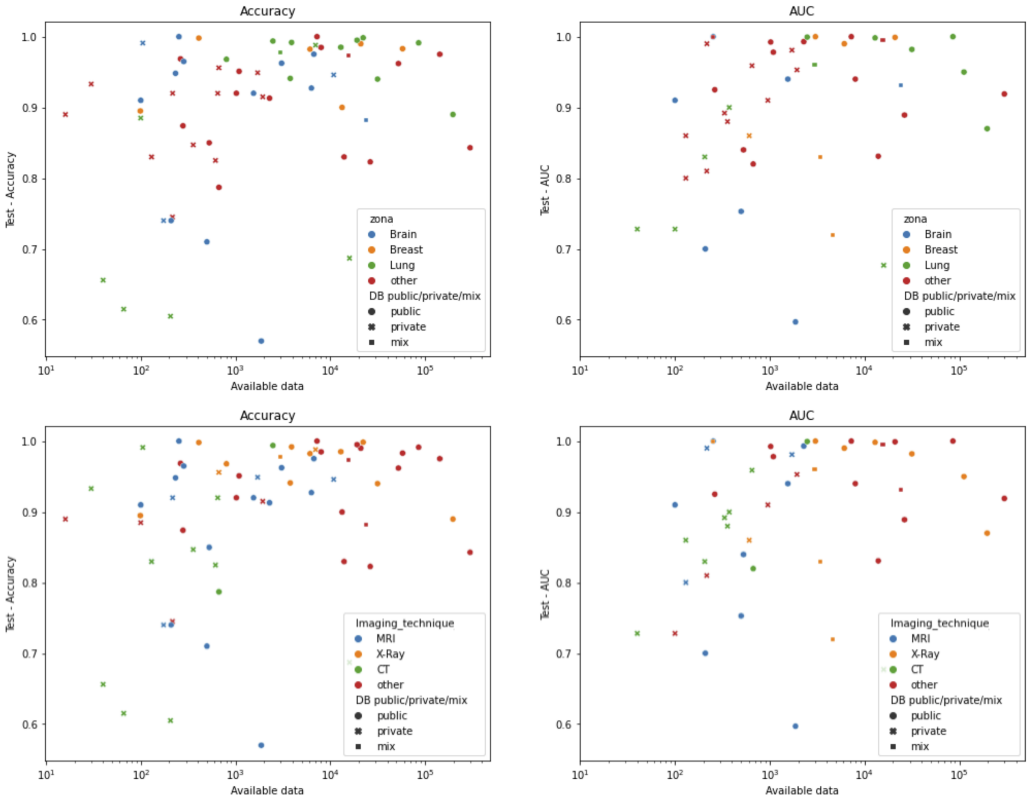
3.5 Conclusion

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.

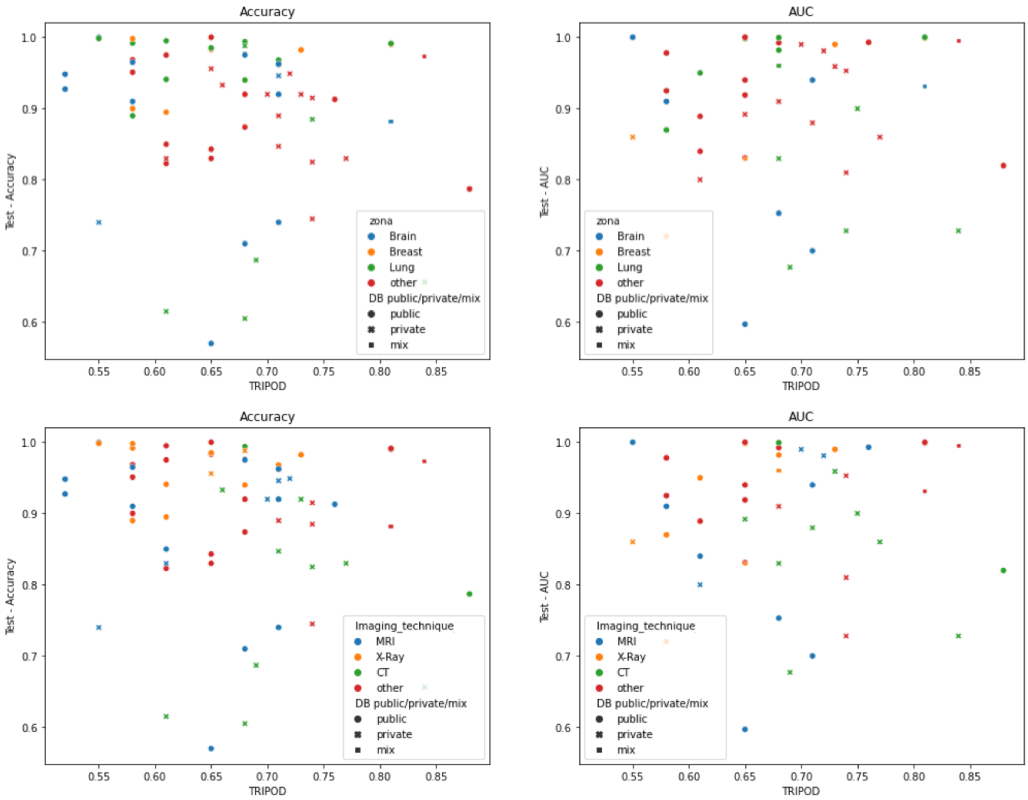
Supplementary Materials



Supplementary Figure 3.1: Quantity of available data with respect to the anatomical region (upper left), the imaging technique (upper right) and the dataset origin (lower left).



Supplementary Figure 3.2: Performances with respect to the available data by anatomical region, imaging technique and dataset origin.



Supplementary Figure 3.3: Performances with respect to the quality of the data (TRIPOD index) by anatomical region, imaging technique and dataset origin.

Chapter 4

Supervised & Unsupervised Clustering

Radiomic and dosiomic-based clustering development for radio-induced neurotoxicity in pediatric medulloblastoma

Stefano Piffer [1,2], Daniela Greto [3], Marzia Mortilla [4], Antonio Ciccarone [5], Cinzia Talamonti [1,2]

[1] National Institute of Nuclear Physics (INFN), Florence Division, Florence, Italy
[2] University of Florence, Florence, Italy
[3] Radiation Oncology Unit, Careggi University Hospital, Florence, Italy
[4] Anna Meyer Children's University Hospital, Radiology Unit, Florence, Italy
[5] Anna Meyer Children's University Hospital, Medical Physics Unit, Florence, Italy

I contributed to think and design the experiment. I personally created and curated the entire database from scratch and collaborated with the expert radiologist and radiotherapist (100%). I performed data preprocessing, features extraction and implemented the machine learning model. I also drafted the article and prepared 100% of the figures.

Abstract

Background: Texture analysis can extract many quantitative image features, offering a valuable, cost-effective and non-invasive approach for individual medicine. Furthermore, multimodal machine-learning procedures could have a large impact for precision medicine, as texture biomarkers can underlying tissue microstructure. Proposed study aims to apply a radiomic and dosiomic analysis on Magnetic Resonance (MR) and dose images, to investigate imaging-based biomarkers of radio-induced neurotoxicity in pediatric patients affected by medulloblastoma.

Methods: In this single-center analysis, 48 children with medulloblastoma treated between 2011 and 2019 were retrospectively enrolled. There were 29 men and 19 women with a mean age of 12 ± 6 years (range: $2 - 23$ years). For each patient, a total of 332 radiomic and dosiomic features were extracted from the region of tumor on the $T1$, $T2$, $FLAIR$ MRI-maps and on radiotherapy dose distribution. Different machine-learning feature selection and reduction approaches were performed to build supervised and unsupervised hierarchical clustering. Moreover, external cluster validation method was applied to get the prediction accuracy.

Results: A greater level of abstraction of input data by combining selection of the most performing features and reduction of dimensionality returns the best performance. The resulting 1-components radiomics signature for clustering, obtained projecting the 4-best selected features, yielded an accuracy of 0.73 with sensitivity, specificity and precision of 0.83, 0.64 and 0.68, respectively.

Conclusion: Machine-learning radiomic-dosiomic approach showed satisfactory stratification performance for unsupervised clustering of pediatric medulloblastoma patient who have experienced radio-induced neurotoxicity. The strategy needs further validation in an external dataset for its potential clinical use in ab initio management paradigms of medulloblastoma.

Keywords: *small data, radiomics, dosiomics, pediatric medulloblastoma, clustering, neurotoxicity*

4.1 Introduction

Medulloblastoma (MB) is the most common brain malignancy in pediatric patients, which accounts for $20-25\%$ of pediatric central nervous system neoplasms (Von Bueren et al., 2016; Ramaswamy and Taylor, 2017). Despite the increase in survival rates in recent years, prognosis of MB patients remains relatively poor, and it strongly depends on clinical and molecular risk factors (Millard and De Braganca, 2016).

In the past the risk stratification was based on age at diagnosis, disease dissemination and extent of resection. Recently a new proposed classification identifies four risk categories (low, standard, high and very high risk) taking into account metastatic stage and genetic and cytogenetic aberrations characterized by very different clinical outcomes and treatment resistance (Archer et al., 2017). MB is currently treated with surgery, chemotherapy and Cranio-Spinal Irradiation (CSI). Cure intensification is based on risk stratification and despite this multimodal approach, about 30% of high-risk patients experience disease relapse (Bouffet, 2021).

Moreover, due to the aggressive therapies and the young age of MB patients, early and late sequelae such as ototoxicity, cardiotoxicity, lung toxicity, neurotoxicity, endocrine deficiency as well as neurocognitive deficits could often develop (Parsons et al., 2011; Tamayo et al., 2011; Packer et al., 2013). In particular neurotoxicity could compromise quality of life in pediatric patients, for example long term neurological sequelae imply that children treated with high dose chemotherapy and/or Radiotherapy (RT) for central nervous system tumors had lower educational outcomes (Lorenzi et al., 2009). The factors that concur to develop neurotoxicity in pediatric patients are argument of scientific discussion; in a recent retrospective review of 113 patients treated with CSI for medulloblastoma, the authors showed a dose response relationship between radiotherapy and neurocognitive impairment (Moxon-Emre et al., 2014). New RT technique, smaller RT field and lower dose are investigated to reduce the impact of radiotherapy on neurotoxicity in central nervous system tumors in children (Seidel et al., 2021). Furthermore, the improvement of diagnostic imaging led to Magnetic Resonance Imaging (MRI) becoming the gold standard in central nervous system tumors (Perreault et al., 2014). Due to the high resolution of morphologic images, MRI guides the clinician with the differential diagnosis and consequently the first approach to the therapeutic path, moreover multiparametric MRI is useful to define treatment response not only detecting tumor shrinkage but also to distinguish pseudo progression and early signs of neurotoxicity (Nichelli and Casagranda, 2021).

A revolutionary approach to medical imaging has been done with radiomics.

The imaging analysis allows the extraction of quantitative features that could be used for clinical purposes. Radiomics derived data when used in combination with clinical data could offer information not only about cancer genotype but also clinical outcome and toxicity treatment correlated (Lambin, 2017; Zhang et al., 2018; Sun et al., 2018; Kickingereder et al., 2016). We hypothesize that non-invasive biomarkers offer great potential for improving stratification in pediatric medulloblastoma. The aim of this study is to analyze MRI features of MB patients treated with surgery, chemotherapy and CSI and look for quantitative features that correlate with clinical outcome. Moreover, the correlation between MRI radiomics features and dosimetric distribution on planning Computed Tomography (CT) are investigated to predict radio-induced neurotoxicity. This toxicity has been identified with radio-necrosis, a condition characterized by the death of tissue due to exposure to high doses of radiation and frequently occurs in the brain. The combination of radiomics and dosiomics (i.e. to extract texture features from dose distribution (Buizza et al., 2021)) analysis is likely to provide non-invasive imaging biomarker.

4.2 Materials and Methods

Dataset

All procedures performed in this study involving human participants were in accordance with the ethical standards of the institution and the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Pediatric Ethics Committee. Since this study is a retrospective analysis and the patients have been anonymously processed, the need for informed consent was waived.

In this single-center analysis, data of patients referred for adjuvant radiotherapy for pathologically confirmed primary MB patients from September 2011 to November 2019 were initially analyzed for further inclusion. The inclusion criteria were (i) availability of postoperative MRI with diagnostic-quality performed after adjuvant RT throughout the follow-up period, (ii) availability of multi-parametric MRI, including axial $T1$, $T2$, and $FLAIR$ maps, (iii) availability of radiotherapy CT, structures set, plan and $3D$ dose volume. Patients with incomplete clinical data, poor tumor tissue quality, and incomplete or poor-quality MR images were excluded from the research. Baseline demographic clinical information including age, gender, metastasis, histologic subtype, and adjuvant therapies (radiation alone, chemotherapy alone or both of them) were collected from the medical record system. Follow-up data were acquired by medical records. The study population included 48 patients with a mean age of 12 ± 6 years, range $2 - 23$ years, 29 men and 19 women.

Regarding the follow up data, the protocol requires MRIs to be repeated every 3 months for the first 3 years after surgery and then every 6 months. This has allowed clinicians to identify whether or not the radio-induced neurotoxicity has occurred and to obtain the ground truth label, hereinafter referred to as 'relapse'. While for RT-treatment, patients treated with Medulloblastoma received either standard-dose (i.e. 30.6 to 39.4 Gy) or reduced-dose (i.e. 18 to 23.4 Gy) radiation to the entire brain and spine. In some cases, patients received a boost to the entire posterior fossa or a focal conformal boost to tumor bed if a residual disease is present; in both cases, total boost volume dose was 45 to 55.4 Gy. Moreover, the radiation was delivered with helical TomoTherapy (Mackie et al., 1993).

Radiomic features extraction

Prior to feature extraction, two fundamental data pre-processing steps was carried out across all the patients: resolution adjustment and images co-registration. Three-dimensional tumor contours were obtained free from radiotherapy process by co-registration of MR images on the centering CTs. Pixels included by the defined tumor contour were applied for feature extraction using PyRadiomics (v2.2.0) (Van Griethuysen et al., 2017), an open-source Python tool. A detailed description of the implementation of these steps and radiomic features extracted by the software is available in the official documentation (`https://pyradiomics.readthedocs.io/en/latest/features.html`). A diagram illustrating images processing and the overall workflow is displayed in Figure 4.1.

Features quantifying tumor phenotypic characteristics on MR and dose images could be grouped as tumor intensity and texture features. In the first category, tumor intensity information are quantified using first-order statistics, obtained from the histogram of entire tumor voxel intensity values. While the second category consists of three-dimensional texture features that are able to quantify the intra-tumoral heterogeneity within a full tumor volume. Textural features were computed based on Gray Level Cooccurrence Matrix ($GLCM$), Gray Level Run Length Matrix ($GLRLM$), Gray Level Size Zone Matrix ($GLSZM$), Gray Level Dependence Matrix ($GLDM$) and Neighbouring Gray Tone Difference Matrix ($NGTDM$).

Features selection & classifier

Data mining and machine learning analysis were performed in the Colab environment (`https://colab.research.google.com`).

To reduce the batch effects, in the feature analysis, the quantitative radiomics raw data were normalized across all patients. Two different feature selection and ranking

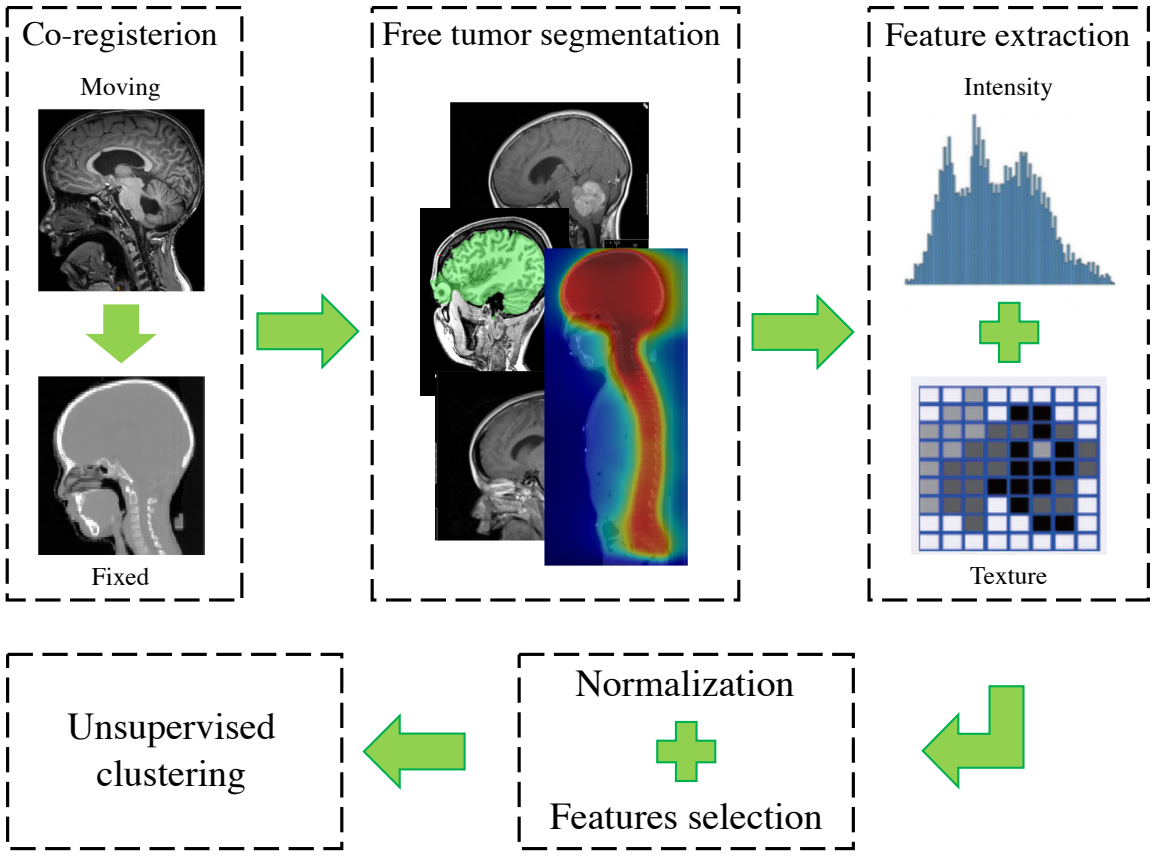


Figure 4.1: Radiomics workflow pipeline.

methods were employed in the analysis based on multivariate filter approaches and on recursive feature elimination. Filter methods are feature-ranking methods, which rank the features using a scoring criterion, and multivariate methods investigate the multivariate interaction within the features and the scoring criterion is a weighted sum of feature relevancy and redundancy. Feature relevancy is a measure of feature's association with the target/outcome variable, whereas feature redundancy is the amount of redundancy present in a particular feature with respect to the set of already selected features (Kohavi and John, 1997). The second approach concerns Recursive Feature Elimination (RFE). It aims to identify the most relevant features from a given dataset by iteratively eliminating less important features based on their contribution to a model's performance. Given an external estimator that assigns weights to features, the algorithm recursively eliminates least important features considering smaller and smaller sets of features. The number of features to eliminate at each iteration is a parameter that needs to be specified. After removing the least important features, the model is retrained on the reduced feature set. That procedure is recursively repeated on the pruned set until a predetermined number

of features remain or until a specific stopping criterion is met. We exploited Mutual Information (MI) and RFE in a 5-fold cross-validation fashion for feature selection and ranking, and Principal Component Analysis (PCA) as dimensionality reduction.

We implemented three different classification methods, Random Forest (RF) - Extreme Gradient Boosting (XGB) - Hierarchical Clustering (HC) and external cluster validation method was applied to get the prediction accuracy. We want to spend a few words about the last two less common classifiers and deepen these concepts of data mining.

XGB is designed to solve supervised learning problems and it is an enhanced version of the traditional gradient boosting algorithm. An ensemble model combines the outputs of multiple weak prediction models to create a stronger and more accurate model. The RF is a popular ensemble that takes the average of many decision trees via bagging. Bagging is short for "bootstrap aggregation", meaning that samples are chosen with replacement (bootstrapping), and combined (aggregated) by taking their average. Boosting is a strong alternative to bagging. Instead of aggregating predictions, boosters turn weak learners into strong learners by focusing on where the individual models went wrong. In Gradient Boosting, individual models train upon the residuals which are the difference between the prediction and the actual results. Instead of aggregating trees, gradient boosted trees learn from errors during each boosting round. The key idea behind XGB is to optimize a specific loss function by iteratively adding weak models and updating the model's predictions based on the residuals. The "eXtreme" refers to speed enhancements since it supports parallel computing. In addition, XGB includes a unique split-finding algorithm to optimize trees, along with built-in regularization to prevent overfitting and improve generalization and which controls the complexity of the model.

Hierarchical cluster analysis is an unsupervised clustering algorithm. The algorithm groups similar objects into groups called clusters. The endpoint is a set of clusters or groups, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. Clustering technique is based on measures of similarity between pair of items in the data set. This similarity is conceived in terms of distance in a multidimensional space, such as the Euclidean distance. Clustering algorithm then group the elements on the basis of their mutual distance, specifically it works out which observations to group based on reducing the sum of squared distances of each observation from the average observation in a cluster. Therefore, whether or not the elements belong to a set depends on how far the element under consideration is from the set itself. The main advantage of hierarchical clustering is that the number of clusters does not have to be defined a priori. Moreover, this technique can be displayed in an attractive, tree-based representation

of the observations, called a dendrogram. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. The distance between data points represents dissimilarities, while height of the blocks represents the distance between clusters.

Concerning cluster validation, external clustering validity approach uses prior knowledge and consists in comparing the results of a cluster analysis to an externally known result, such as externally provided class labels. It measures the extent to which cluster labels match pre-existing clustering structure (reference labels). We preferred this method since we know the "true" cluster number and reference labels in advance (Rendón et al., 2011). Measures of the machine learning classifier performance included: accuracy, sensibility, sensitivity (recall), precision, *F*-score and MCC.[1].

4.3 Results

A total of 332 radiomic and dosiomic features were extracted from each patient, 83 for each of the four available images series (three MRI sequences and dose distribution). Their pairwise correlation cluster map can be found in Supplementary File 4.1. Among these, feature selection worked by selecting the k best most informative features based on MI statistical test. In our case, the 20 best features derived from dose, $T1$ and $T2$ weighted images are made explicit in Table 4.1. Their univariate and bivariate distribution in our population based on relapse occurrence can be found in Supplementary File 4.2.

The first step of our strategy for features selection was to consider the best 20 features (correlation matrix can be find in Supplementary File 4.3) based on the explained variance; in fact in Figure 4.2 it can be seen how already with only 20 components (intended as the number of features) it is possible to maintain as much as 95% of the variability present in the data. A higher explained variance indicates a better fit and suggests that the model is capturing a significant portion of the underlying relationships between the variables. Taking into account the second selection and ranking method, in RFE we set the achievement of 20 features as a stopping criterion following what we learned a little while ago. The process was repeated with four common external estimators (Logistic Regression (LG), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB)) and the selected features were compared with those identified by MI statistical test. From the histogram in Figure 4.3 it is possible to notice how certain variables are more frequently present in the subsets of 20 features and are also the same ones that are found in the first

[1]MCC is a measure of the quality of binary classifications in machine learning, ranging from $+1$ (perfect prediction) to 0 (average random prediction) and -1 (inverse prediction).

Table 4.1: Characteristics of each selected feature and relative class according to PyRadiomics official documentation (`https://pyradiomics.readthedocs.io/en/latest/features.html`).

Sequence	Features	Features class	Acronym
Dose	Small Area High Gray Level Emphasis	Gray Level Size Zone Matrix	SAHGLE_glszm_Dose
	Zone Entropy		ZE_glszm_Dose
	Inverse Difference Moment Normalized	Gray Level Co-occurrence Matrix	IDMN_glcm_Dose
	Busyness	Neighboring Gray Tone Difference Matrix	Busyness_ngtdm_Dose
	Small Dependence High Gray Level Emphasis	Gray Level Dependence Matrix	SDHGLE_gldm_Dose
T1w	Maximal Correlation Coefficient	Gray Level Co-occurrence Matrix	MCC_glcm_T1
	Sum Square		SS_glcm_T1
	Gray Level Variance	Gray Level Dependence Matrix	GLV_gldm_T1
	Gray Level Variance	Gray Level Run Length Matrix	GLV_glrlm_T1
	Coarseness	Neighboring Gray Tone Difference Matrix	Coarseness_ngtdm_T1
T2w	Gray Level Variance	Gray Level Dependence Matrix	GLV_gldm_T2
	Difference Average	Gray Level Co-occurrence Matrix	DV_glcm_T2
	Difference Entropy		DE_glcm_T2
	Difference Variance		DA_glcm_T2
	Contrast		Contrast_glcm_T2
	Cluster Tendency		CT_glcm_T2
	Sum Squares		SS_glcm_T2
	Gray Level Non-Uniformity Normalized	Gray Level Run Length Matrix	GLNUN_glrlm_T2
	Gray Level Variance		GLV_glrlm_T2
FLAIR	Busyness	Neighboring Gray Tone Difference Matrix	Busyness _ngtdm_FLAIR

places of the ranking proposed by the mutual information analysis. As second step, features reduction was conducted exploiting the PCA technique, which permits a dimensionality reduction. It combines input data by projecting them into a lower number of components, four and one in our case, following the rule of thumb to select one feature every 10/15 variables. Thereby we increase the informative power of the remaining features, but we cannot have a direct definition of what each single feature describes.

The evaluation metrics according to the various strategies for features selection and features reduction are shown in the Table 4.2. Considering the best performing strategy, its accuracy is 0.73 with 35/48 correctly classified patients. In particular,

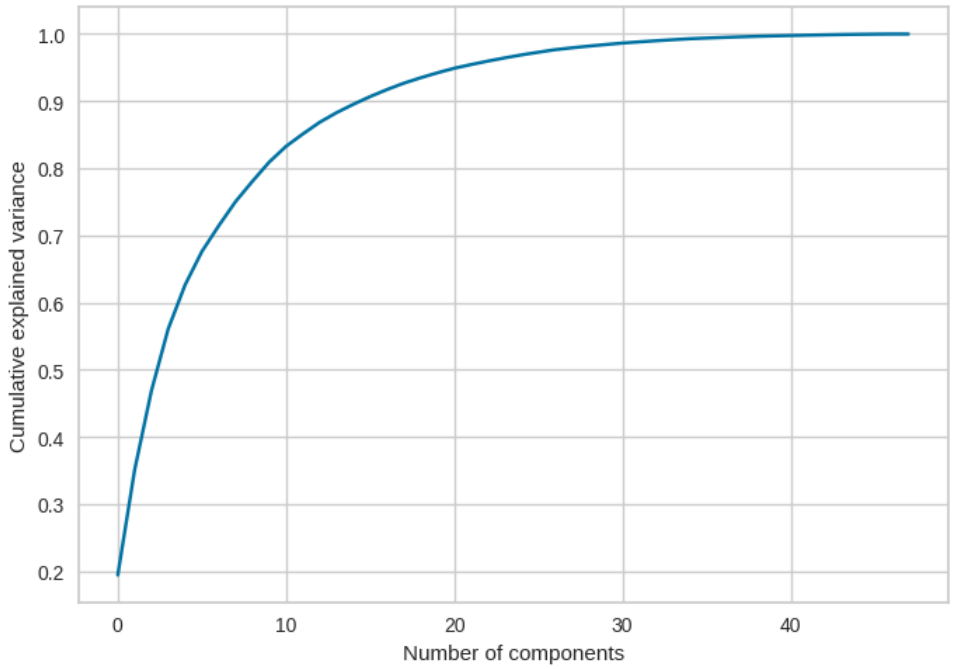


Figure 4.2: Cumulative explained variance as function of the number of features. Already with 20 features it is possible to account for 95% of the total variance.

analyzing Sensitivity and Specificity the model demonstrates good prediction power at identifying patients who have suffered radio-induced toxicity.

4.4 Discussion

In recent years, an increasing number of reports demonstrated the added value of machine learning-based radiomics analysis to clinical and conventional MRI characteristics in pediatric MB, pointing out the potential to predict molecular markers and molecular subtype, to improve survival prediction, to evaluate the intratumoral heterogeneity and to boost prognostic models (Liu et al., 2022; Yan et al., 2020; Zheng et al., 2021; Zhou et al., 2021). However, to our knowledge, the relationships between the combination of radiomic-dosiomc features and radio-induced neurotoxicity of MB patients has not been investigated. The main finding of this study is that our machine learning approach showed satisfactory stratification performance for clustering of pediatric medulloblastoma patient who have experienced radio-induced neurotoxicity based on radiomic and dosiomic features extracted from MR and dose images. The accurate stratification of pediatric medulloblastoma patient is highly desired to select the most appropriate treatment (Liu et al., 2022), especially in view of a dose de-escalation with the same disease control. Indeed, patients treated with higher doses are prone to experienced intellectual declines and presence of radio-

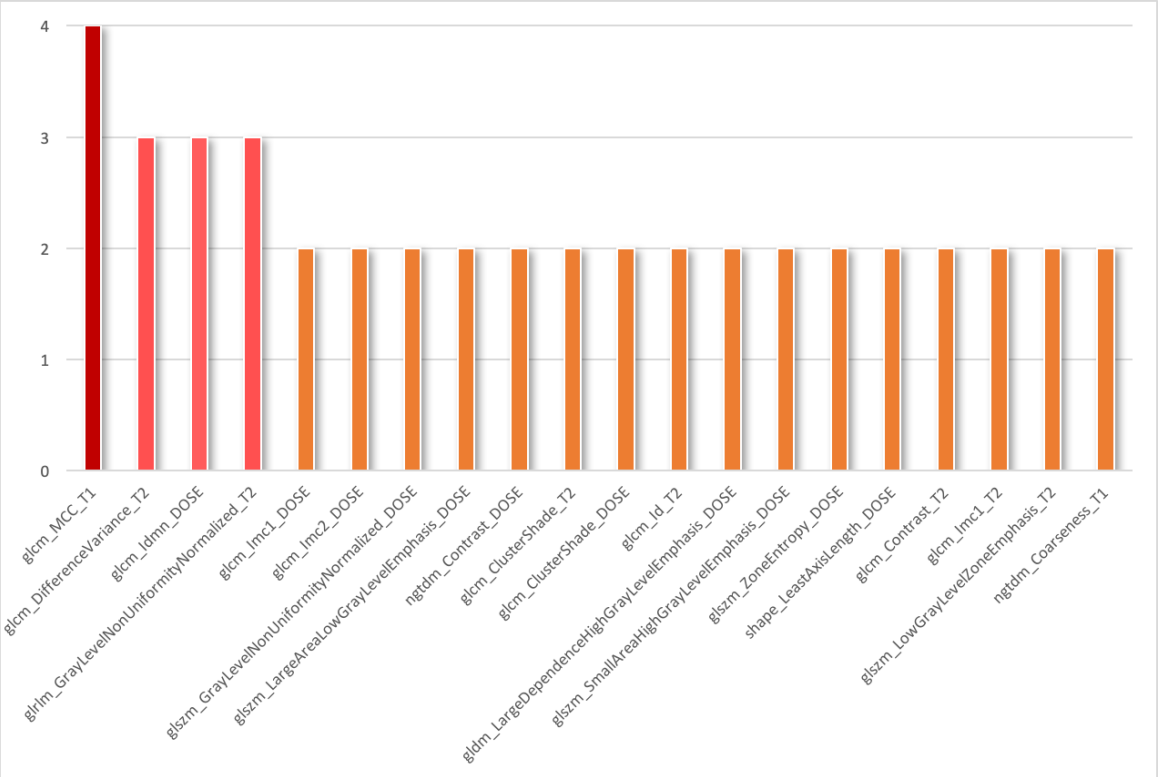


Figure 4.3: Histogram of the frequency of the top features identified with the different selection methods.

induced neurotoxicity are associated with worse intellectual outcome (Moxon-Emre et al., 2014).

The machine learning protocol followed in this study foresees to examine the dose distribution calculated for the RT treatment and the MR images of the first follow-up after radiotherapy. From the quantitative data extracted from these images it was possible to establish a radiomic signature that has the potential to highlight patients in whom radio-induced damage will develop. This could have a great clinical response, because it gives the physician the possibility to intervene promptly with adequate therapies and reduce complications, since the detriments caused by ionizing radiations have a medium-to-long latency.

Considering the features extraction and reduction strategies, it was possible to appreciate that a greater level of abstraction of input data by combining the selection of the most performing features and the reduction of dimensionality with PCA returns a better prediction performance. Combining the 20-best features by projecting them into a lower number of components (four or just one dimension) have increased the informative power of the input data, compared to directly considering the 20-best features. The best result was obtained by taking into consideration the 4-best features according to the ranking given by the mutual information test and projecting these

Table 4.2: Classifier evaluation metrics; MCC, Matthews Correlation Coefficient.

	Explained Variance	Model	Accuracy	Sensitivity	Specificity	Precision	F-score	MCC
All	1.00	HC	0.56	0.16	1.00	1.00	0.28	0.29
		RF	0.58	0.64	0.52	0.59	0.62	0.16
		XGB	0.48	0.58	0.38	0.48	0.53	-0.04
20-best	0.95	HC	0.60	0.56	0.65	0.64	0.60	0.21
		RF	0.65	0.80	0.48	0.63	0.70	0.29
		XGB	0.63	0.68	0.57	0.63	0.65	0.25
4-best	0.67	HC	0.65	0.36	0.96	0.90	0.51	0.39
		RF	0.69	0.88	0.48	0.65	0.75	0.39
		XGB	0.65	0.68	0.61	0.65	0.67	0.29
20-best → 4-PCA	0.91	HC	0.63	0.54	0.71	0.65	0.59	0.25
		RF	0.67	0.81	0.50	0.66	0.72	0.32
		XGB	0.67	0.72	0.61	0.67	0.69	0.33
20-best → 1-PCA	0.74	HC	0.63	0.64	0.62	0.58	0.61	0.25
		RF	0.65	0.80	0.48	0.63	0.70	0.29
		XGB	0.65	0.60	0.70	0.68	0.64	0.30
4-best → 1-PCA	0.67	HC	0.71	0.60	0.83	0.79	0.68	0.44
		RF	***0.73***	***0.83***	***0.64***	***0.68***	***0.75***	***0.47***
		XGB	0.67	0.61	0.72	0.67	0.64	0.33
20-RFE - LR	0.93	HC	0.63	0.40	0.87	0.77	0.53	0.30
		RF	0.60	0.68	0.52	0.61	0.64	0.20
		XGB	0.60	0.64	0.57	0.62	0.63	0.21
20-RFE - DT	0.89	HC	0.58	0.24	0.96	0.86	0.38	0.28
		RF	0.60	0.76	0.43	0.59	0.67	0.21
		XGB	0.58	0.68	0.48	0.59	0.63	0.16
20-RFE - RF	0.91	HC	0.42	0.32	0.52	0.42	0.36	-0.16
		RF	0.69	0.88	0.48	0.65	0.75	0.39
		XGB	0.71	0.76	0.65	0.70	0.73	0.42
20-RFE - GB	0.92	HC	0.54	0.60	0.48	0.56	0.58	0.08
		RF	0.67	0.80	0.52	0.65	0.71	0.34
		XGB	0.69	0.64	0.74	0.73	0.68	0.38

four variables into a single component. In this scenario, satisfactory results of the various metrics were obtained for all three classifiers; in particular the podium was awarded by the RF algorithm with an accuracy of 0.73 and an MCC of 0.47. This outcome is remarkable given the small number of the database and indicates a good agreement between the predicted and actual classifications; probably also due to the simplicity of the trained model which made it possible to contain overfitting. The resulting drawback of this approach is that we no longer have a direct definition of what each single feature describes. Taking a step back and examining the description of the 4-best identified features, it can be seen how they take into account small size zones with high gray-level values for the dose distribution indicating the dose

hot-spots, disparity in intensity values among neighboring voxels and measures of heterogeneity on differing intensity levels pairs that deviate more from the mean for $T2$-weighted and complexity of the texture for $T1$-weighted maps. All of them can be seen as describing two fundamental properties: homogeneity and heterogeneity of the underlying tissue and dose microstructure (Zheng et al., 2022; Chang et al., 2021). Further, it must be highlighted that features extracted from dose images also contribute to the construction of the radiomic signature. In addition, following the scores presented in Figure 4.3, it is confirmed that certain features are robust with respect to the various feature selection methods. In particular, by comparing the ranking given by the statistical test of the MI with the RFE for the various external estimator, it is clearly seen that in addition to the 4-best features previously indicated, the various subsets also include features that concern heterogeneity in texture patterns, spatial rate of change, local homogeneity in the image and similarity of gray-level intensity values, emphasizing our previous assumption.

Taking into account the classification methods, all three techniques showed good results, comparable to each other and without running into overfitting. Between the two supervised algorithms, RF shows on average slightly better performance than XGB, probably due to for its simplicity, scalability, robustness to noise and therefore it is possible to train better even with small data size available. In fact, RF has fewer hyperparameters to tune compared to XGB, which has a wide range of hyperparameters that control the behavior of the boosting process and a proper tuning of hyperparameters is crucial for achieving optimal performance with XGB. On the other hand, from the results in Table 4.2 we can say that unsupervised clustering has intermediate performances respect to the two systems just described. It is necessary to point out that hierarchical clustering does not require any prior assumptions and the classification we found, arose spontaneously from the data without forcing. The hierarchical cluster tree may naturally divide the data into distinct, well-separated clusters. This can be particularly evident in the attractive dendrogram representation created from data where groups of objects are densely packed in certain areas and not in others (Supplementary File 4.4). At the bottom of the tree, each leaf represents one of 48 observations in the data set. As we move up the tree, the leaves which are similar to each other are grouped into branches. The branches then define combinations of observations until the top, where there is only one root. The height of the dendrogram indicates the similarity between observations or clusters of observations. Two observations meeting at a lower branch will be more similar to each other than another pair of observations meeting at a higher branch.

To sum up, we obtained comparable performance applying two intrinsically dif-

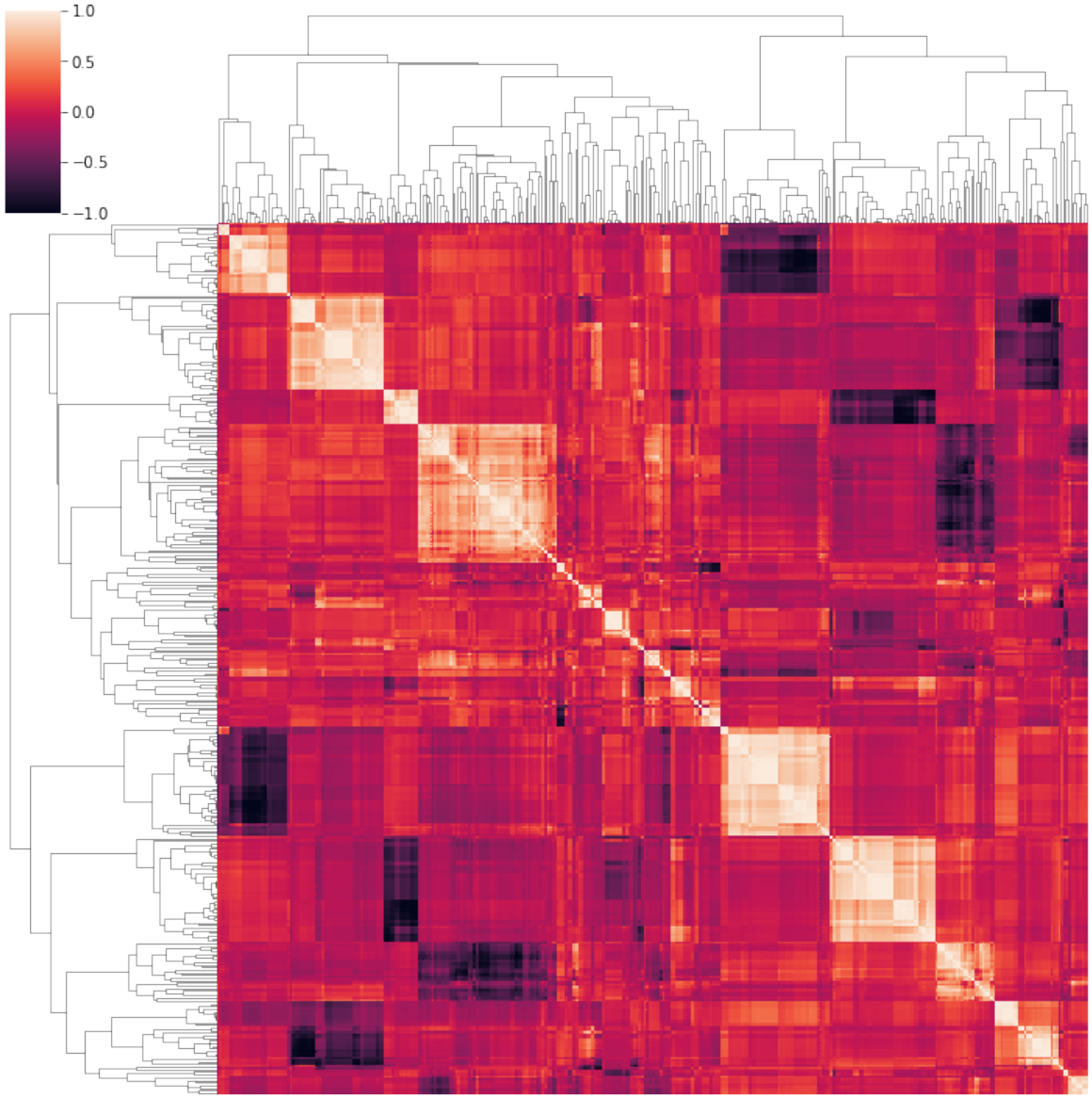
ferent methods; on one hand supervised learning algorithms learn patterns and relationships between features and target variable, on the other unsupervised learning algorithm groups similar data points into clusters based on their distances or similarities discovering inherent patterns without any predefined target variable. This indicates the goodness of the data available and the care taken in creating the database, albeit of modest dimensions.

Nonetheless, some limitations of this study need to be addressed. First, pediatric MB is a rare tumor and although our research extends over 8 years, the patients' cohort is quite limited. In addition, the data were all from a single institution although this peculiarity has allowed us to build a homogeneous, complete and balanced database. Second, an external patient population for assessing of the radiomics signature generalizability is not available. Future investigations will require data exchange between different institutions to obtain higher volume database thanks to which it could be possible obtain performance more reflective of the real predictive power of current method.

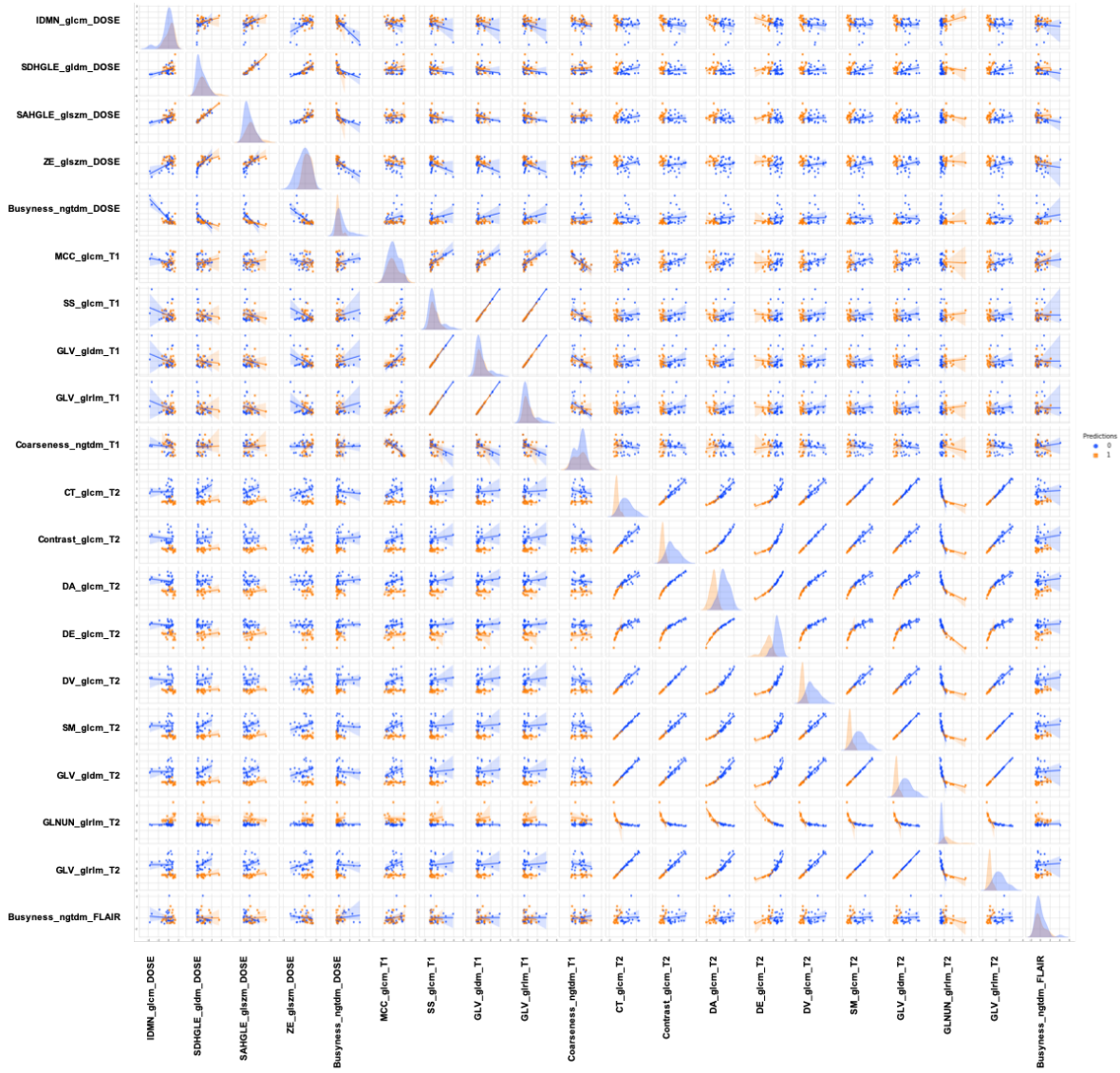
4.5 Conclusion

We believe the current imaging techniques may potentially be further equipped to better classify and safely diagnose possible complications and the current study demonstrated proof-of-concept results for integrating radiomics protocol. In this regard, radiomics and dosiomics may prove a valuable and cost-effective aid by providing non-invasive quantitative data that integrate qualitative image information already available.

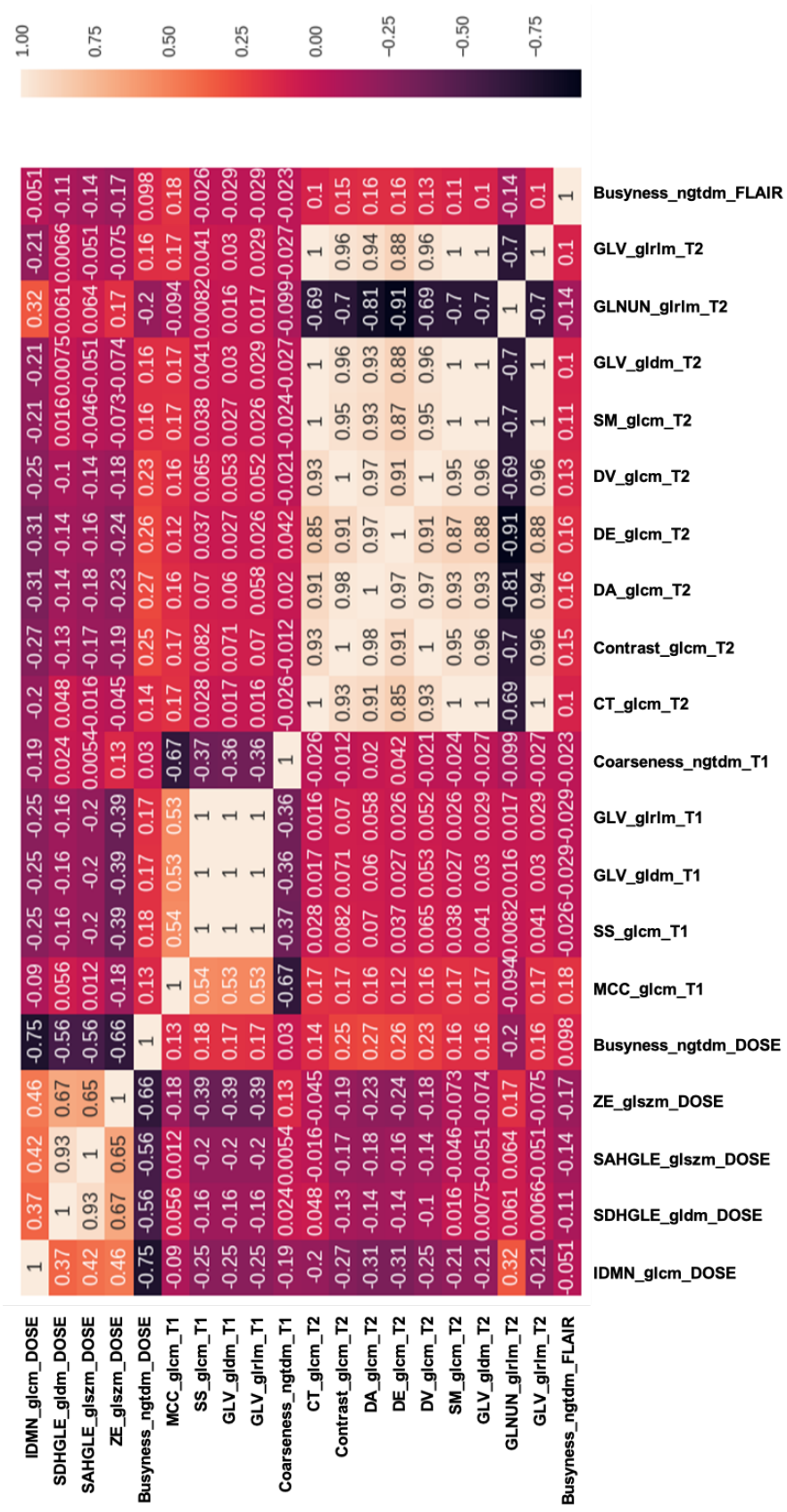
Supplementary Files



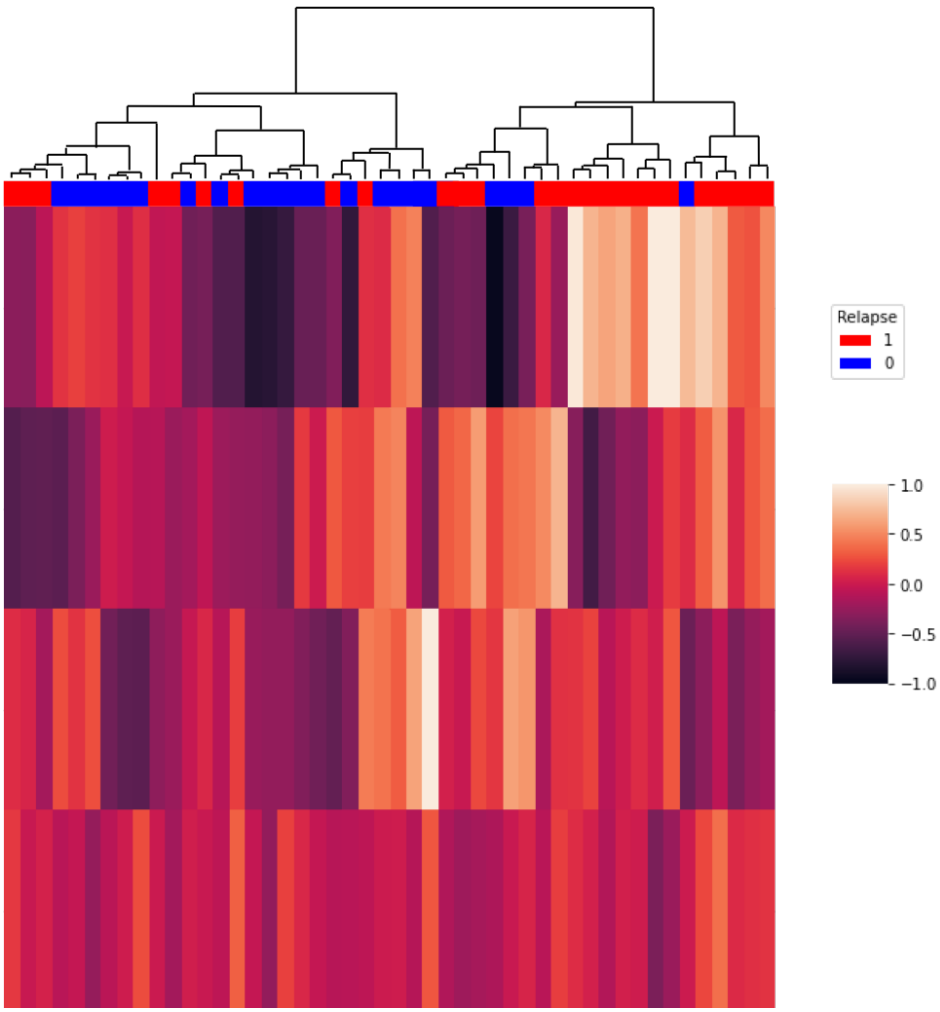
Supplementary Figure 4.1: Pairwise correlation cluster map concerning all extracted features.



Supplementary Figure 4.2: Univariate and bivariate distribution with regression lines for the 20-best selected features in relation to the relapse occurrence.



Supplementary Figure 4.3: Correlation matrix of the 20-best extracted features.



Supplementary Figure 4.4: Heat map of the reduced 4-best radiomics features signature. Hierarchical clustering with dendrogram of relapse occurrence is on the top. The red/blue bar indicates the true labels.

Chapter 5

Aggregation of Public Datasets

Making Data Big for a Deep-learning Analysis: Aggregation of Public COVID-19 Datasets of Lung Computed Tomography Scans

Francesca Lizzi [1,2], Francesca Brero [3,4], Raffaella Fiamma Cabini [3,5], Maria Evelina Fantacci [2,6], Stefano Piffer [7,8], Ian Postuma [3], Lisa Rinaldi [3,4], Alessandra Retico [2]

[1] Scuola Normale Superiore, Pisa, Italy
[2] National Institute of Nuclear Physics (INFN), Pisa Division, Pisa, Italy
[3] National Institute of Nuclear Physics (INFN), Pavia Division, Pavia, Italy
[4] Department of Physics, University of Pavia, Pavia, Italy
[5] Department of Mathematics, University of Pavia, Pavia, Italy
[6] Department of Physics, University of Pisa, Pisa, Italy
[7] National Institute of Nuclear Physics (INFN), Florence Division, Florence, Italy
[8] University of Florence, Florence, Italy

I substantially contributed to develop and implement the in-house lung segmentation algorithm based on active contours in Matlab (80%). I suggested the appropriate windowing range of Hounsfield Units and Dice Similarity Coefficient as evaluation metrics.

Abstract

Lung Computed Tomography (CT) is an imaging technique useful to assess the severity of $COVID-19$ infection in symptomatic patients and to monitor its evolution over time. Lung CT can be analysed with the support of deep learning methods for both aforementioned tasks. We have developed a $U-net$ based algorithm to segment the $COVID-19$ lesions. Unfortunately, public datasets populated with a huge amount of labelled CT scans of patients affected by $COVID-19$ are not available. In this work, we first review all the currently available public datasets of $COVID-19$ CT scans, presenting an extensive description of their characteristics. Then, we describe the design of the $U-net$ we developed for the automated identification of $COVID-19$ lung lesions. Finally, we discuss the results obtained by using the different publicly available datasets. In particular, we trained the $U-net$ on the dataset made available within the *COVID-19 Lung CT Lesion Segmentation Challenge 2020*, and we tested it on data from the *MosMed* and the $COVID-19-CT-Seg$ datasets to explore the transferability of the model and to assess whether the image annotation process affects the detection performances. We evaluated the performance of the system in lesion segmentation in terms of the *Dice index*, which measures the overlap between the ground truth and the predicted masks. The proposed $U-net$ segmentation model reaches a *Dice index* equal to 0.67, 0.42 and 0.58 on the independent validation sets of the *COVID-19 Lung CT Lesion Segmentation Challenge 2020*, on the *MosMed* and on the $COVID-19-CT-Seg$ datasets, respectively. This work focusing on lesion segmentation constitutes a preliminary work for a more accurate analysis of $COVID-19$ lesions, based for example on the extraction and analysis of radiomic features.

Keywords: $COVID-19$, *Lung CT*, $U-net$, *Data Aggregation*, *Image Segmentation*

5.1 Introduction

Lung Computed Tomography (CT) is a very sensitive medical imaging technique to detect lung lesions due to $COVID-19$. It can be used for the diagnosis, prognosis and for monitoring the disease evolution over time. Despite the use of CT for diagnosis is not recommended by the World Health Organization (Organization, 2020), lung CT analysis can be very informative regarding the severity of the disease and its time evolution (Fang et al., 2021). The use of CT in clinical practice for $COVID-19$ diagnosis in symptomatic patients has been explored. Since the unexpected outbreak of the pandemic, physicians tried to use CT imaging of the chest to diagnose $COVID-19$ disease. The first publication describing in details radiological findings of CT was published in January, the 24 of 2020 (Huang et al., 2020) and it describes the radiological findings of the majority of $COVID-19$ hospitalized patients of this study, such as bilateral multiple lobular and sub-segmental areas of consolidation and bilateral ground-glass opacity. Afterwards, several studies have been published to describe the radiological findings of $COVID-19$ chest CT (Carotti et al., 2020). A summary of all possible findings and their incidence is reported in Table 5.1.

We underline that the dataset used by (Huang et al., 2020) contains a very limited number of CT scans (41 patients) and it is private. Most of the chest CT findings cannot be related exclusively to $COVID-19$ because they are nonspecific signs of disease and they are strongly related to the stage of the disease. This means that there are other forms of pneumonia that may have the same signs such as $SARS-CoV-1$ and $MERS-CoV$. For this reason, the World Health Organization (WHO) defined

Table 5.1: Summary of $COVID-19$ chest CT findings and their incidence on the population. The normal chest CT findings are also associated to symptomaticity (Huang et al., 2020).

Findings	Incidence
Normal chest CT findings	10.6% (95% CI: 7.6%, 13.7%)
Ground-Glass opacity, Lower lobe involvement, Bilateral abnormalities, Vascular enlargement, Posterior predilection,	High incidence (More than 70%)
Consolidation, linear opacity, septal thickening and/or reticulation, crazy-paving pattern, air bronchogram, pleural thickening, halo sign, bronchiectasis, nodules, bronchial wall thickening, reversed halo sign	Intermediate (between 70% and 10%)
Pleural effusion, lymphadenopathy, tree-in-bud sign, central lesion distribution, pericardial effusion, cavitating lung lesions	Low incidence (less than 10%)

as "confirmed case" the patient that have been tested positive for $COVID-19$ $RT-PCR$, irrespective of clinical signs and symptoms (Organization, 2020). Furthermore, it is necessary to differentiate the $COVID-19$ infections not only from other viral pneumonia but also from bacterial pneumonia, such as mycoplasma pneumonia (Ishiguro et al., 2019). The use of chest CT to diagnose $COVID-19$ is, hence, under discussion since it implies the use of ionizing radiation (Adams et al., 2020b) while its ability to monitor the progression of the disease seems to be a promising way to use lung CTs (Adams et al., 2020a). Artificial Intelligence (AI) is a powerful instrument that allows to analyse a huge quantity of data, such as CT scans and, hence, it can be used to monitor and study $COVID-19$ CT signs (Gülbay et al., 2021). Unfortunately, AI implementations require a great amount of data, which may be not easily available. This is especially true when deep-learning methods are used. Since the beginning of the pandemic, some lung CT scans of $COVID-19$ patients have been released by different institutions following different guidelines for both image acquisition and annotation (ground truth). In this work, all the public lung $COVID-19$ CT datasets, to the best of our knowledge, suitable for training AI-based systems are reviewed. In this work, we present an extensive description of the currently publicly available datasets, which present different characteristics, and discuss the segmentation results obtained by using them. In particular, we trained a $U-net$ on the dataset released within the *COVID-19 Lung CT Lesion Segmentation Challenge 2020* (An et al., 2020) and tested it on data from *MosMed* (Morozov et al., 2020a) and $COVID-19-CT-Seg$ datasets (Ma et al., 2020). Finally, the limits and the advantages of aggregating this kind of data are discussed.

5.2 AI and Medical Image Dataset Issues

AI has been used to analyse and process CT to diagnose $COVID-19$, to segment lesions inside the lungs and, also, in longitudinal studies to track the evolution of the disease (Ma et al., 2020). AI based methods, especially deep-learning ones, need a huge amount of labelled data that are not easy to collect and share. As already described in the introduction, some studies use private datasets which do not allow a fair comparison with other AI based systems. Furthermore, the characteristics of CT images depend on the scanner, on the acquisition and the reconstruction protocols and on other information which may not be available. This can be due also to the anonymization process needed to preserve subjects' privacy or to the use of image format different from DICOM, such as the NIfTI format. DICOM is the most used image format for medical images and it contains several metadata in its header. The DICOM header stores many information, some of which is Protected Health

Information (PHI) or private keys that are inserted and encoded by the manufacturer and may contain PHI as well. On the other hand, some metadata, such as anode characteristics or X-ray parameters, do not contain PHI and they can be useful in analysing images. For all these reasons, anonymizing a DICOM file is not a trivial problem and dataset may include images in a different format such as NIfTI (Moore et al., 2015). Deep learning based methods often require the association with a label depending on the task we want to solve. Many approaches are based on supervised learning and, hence, image annotation plays a crucial role. Usually, medical image labels are given by one or more radiologists with experience in the specific field, and image annotation is a very time-consuming task. This is the reason why there is a general lack of public labelled datasets of medical images. In order to save time, it may happen that the labelling is made with the support of an automatic tool and then labels are adjusted manually by one or more physicians.

5.3 Lung CT Datasets

In this section, the currently available datasets of lung CT and their annotation process are reported. The dataset are: *COVID-19 Lung CT Lesion Segmentation Challenge 2020* dataset, *MosMed* dataset, $COVID-19-CT-Seg$ dataset and $TCIA-COVID-19-AR$.

COVID-19 Lung CT Lesion Segmentation Challenge 2020 Dataset

The *COVID-19 Lung CT Lesion Segmentation Challenge 2020* (Challenge dataset) dataset is a public dataset made by 199 unenhanced chest CT with positive $RT-PCR$ for $SARS-CoV-2$ patients (An et al., 2020), published as training set in the occasion of the *COVID GrandChallenge* (`https://covidsegmentation.grand-challenge.org`). Each CT is annotated voxel-wise and indicates all the $COVID-19$ lesions in a unique mask. Data has been provided by The Multi-national NIH Consortium for CT AI in $COVID-19$ via the NCI TCIA public website in Neuroimaging Informatics Technology Initiative (NIfTI) format. Annotations have been made using a $COVID-19$ segmentation model provided by NVIDIA that takes a full CT chest volume and produces pixel wise segmentation masks of $COVID-19$ lesions. These segmentation masks have been adjusted manually by a board of certified radiologists in order to give $3D$ consistency to the lesion masks. The annotations of the training set have been published in the context of the challenge while the system performance has been evaluated by challenge organizers on an independent validation set of 50 CT scans, for which the lesion annotations were not publicly released. A third set, an

independent test set consisting of 46 CT scans, was used to the define the final ranking among the participants, and, also in this case, the lesion segmentation annotations were not publicly released.

MosMed Dataset

MosMed (Morozov et al., 2020a) is a dataset of $COVID-19$ Chest CT scans collected by the Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department. It includes 1110 CT studies taken from 1110 patients and it is provided with a labelling that consists of 5 classes, based on the percentage of involved lung parenchyma. A small subset of class $CT-1$ cases (50 patients) has been annotated by expert radiologists with the support of MedSeg software (2020 Artificial Intelligence AS). The image annotations consist of binary masks in which white voxels represent both Ground-Glass opacities and consolidation. Both CT scans and annotations were provided in NIfTI format. During the $DICOM-to-NIfTI$ conversion only one every $10th$ image was preserved (MosMed, site).

COVID-19-CT-Seg Dataset

The $COVID-19-CT-Seg$ dataset is a collection of CT scans made available by the Coronacases Initiative and Radiopaedia (Ma et al., 2020) and contains 20 CT scans of patients resulted positive for $RT-PCR\ COVID-19$ infection. It is a public dataset which contains annotations related to both lung and infection localization. The ground truth has been made in three steps: first, junior radiologists ($1-5$ years of experience) delineated the annotations of lungs and infections, then two radiologists ($5-10$ years of experience) refined the labels and finally the annotations were verified and optimized by a senior radiologist (more than 10 years of experience in chest radiology). The annotations have been produced with ITK-SNAP software. Ten cases of this dataset were provided in 8-bit depth which are not commonly used in clinical practice.

TCIA-COVID-19-AR

The $TCIA-COVID-19-AR$ (Desai et al., 2020) is a dataset of $COVID-19$ cases taken from a rural population, which is often underrepresented in public datasets. It contains 24 CT scans of patients with both lung lesions due to $COVID-19$ and control cases. Each patient is described by a set of clinical data correlates that includes key radiology findings. Moreover, for each patient the information about Intensive

Care Unit (ICU) admission is included while annotations on images are not included in this dataset.

5.4 COVID-19 Lesion Segmentation

We developed an automatic system which can segment $COVID-19$ lesions based on a $U-net$ (Ronneberger et al., 2015) in the framework of the $COVID-19$ *Lung CT Lesion Segmentation Challenge 2020* (GrandChallenge, site). First, a bounding box which contains the lungs has been built for each CT scan to reduce as much as possible the background from the images. An in-house lung segmentation algorithm based on active contours was developed for this purpose and implemented in Matlab (The MathWorks, Inc.). This algorithm, which accurately segments the lung parenchyma in absence of lesions, has very limited performance on CT scans of subjects with $COVID-19$ lesions. The CT images have been cropped to the bounding boxes, resized to a matrix of $200 \times 150 \times 100$ voxels and a CT windowing in $[-1000, 300]$ range of Hounsfield Units has been applied on them to enhance the $COVID-19$ lesions. A schematic representation of the used $U-net$ is reported in Figure 5.1.

We trained the network on the Challenge training dataset of 199 CT scans, using a weighted crossentropy as loss function, and we tested it on the Challenge validation set (independent from the training set). In order to have a sufficient number of samples, we applied data augmentation with rotations, zooming and elastic transformation to the training set. We tested the network also on the 50 annotated cases of *MosMed* and on the 10 annotated cases of the $COVID-19-CT-Seg$ dataset. The *MosMed* dataset contains images and labels taken in a very different way with respect to those of the Challenge dataset. The $COVID-19-CT-Seg$ dataset has been built in a more similar way to the Challenge one for both data characteristics, such as slice thickness, and labelling process. We evaluated the segmentation performance of the trained network model in terms of *Dice index* (Equation) defined as:

$$Dice_{metric} = \frac{2 \cdot |M_{true} \cap M_{predict}|}{|M_{true}| + |M_{predict}|} \tag{1}$$

where M_{true} is the ground truth mask and $M_{predict}$ is the predicted one. We participated in the challenge, obtaining a *Dice index* equal to 0.67 on the challenge validation set (GrandChallenge, site). Then, we computed the segmentation performance of the trained model on the *MosMed* dataset obtaining a *Dice* of 0.42, and on the $COVID-19-CT-Seg$ dataset, obtaining a *Dice* of 0.58.

We show in Figure 5.2 a visual comparison between the reference $COVID-19$ lesion masks and the ones predicted by the trained $U-net$ for a representative CT

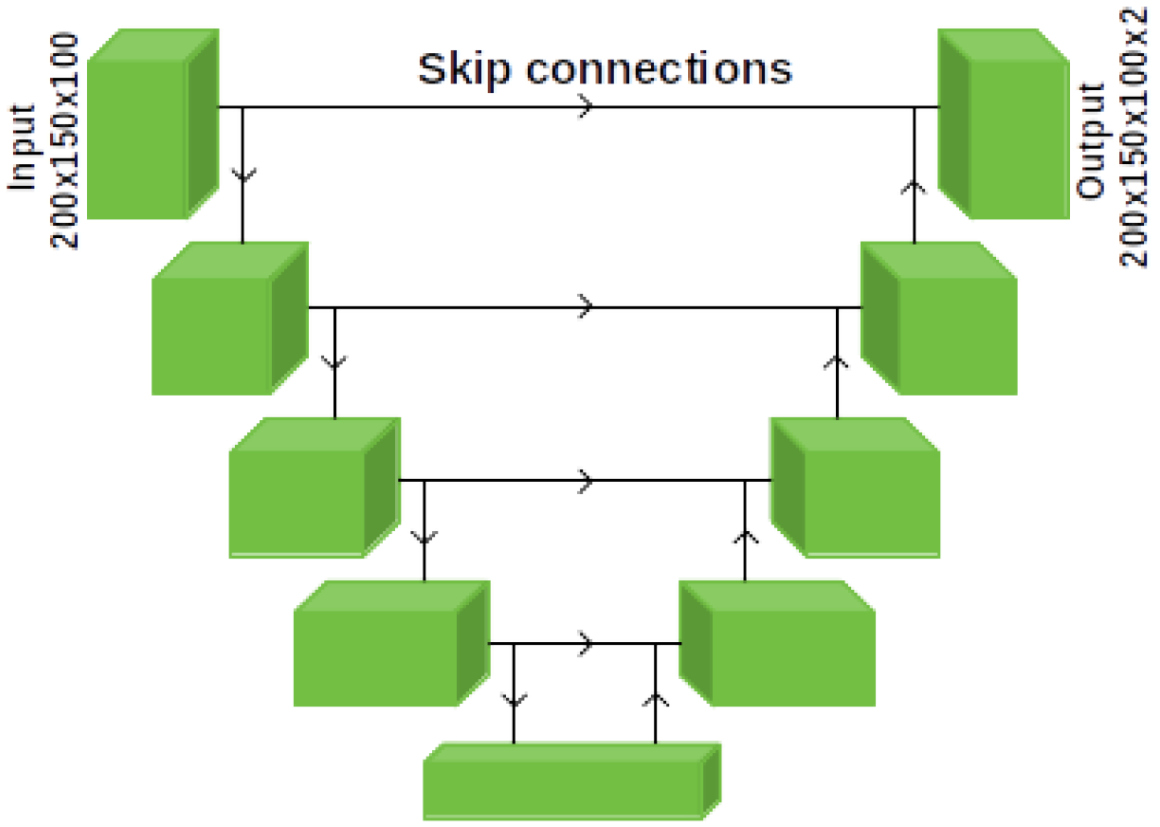


Figure 5.1: $U-net$ summary: the U-shaped neural network is made of 5 levels of depth. In the left path (compression), the input is processed through convolutions, activation layers (ReLu) and instance normalization layers, while in the right one (decompression), in addition to those already mentioned, $3D$ Transpose Convolution (de-convolution) layers are also introduced. Each block (green) is made of 3 convolutional layers.

scan of the *MosMed* and of the $COVID-19-CT-Seg$ dataset.

5.5 Discussion and Conclusions

We obtained good results in terms of the *Dice index* as regards the segmentation of the lung lesions related to $COVID-19$ infection on the Challenge dataset compared to literature (Ma et al., 2020). The results obtained on the other two datasets are not good as the first one. We underline that on the dataset more similar to the Challenge one, the $COVID-19-CT-Seg$ dataset, we obtained better results compared to *MosMed*. As expected, we conclude that aggregating data from different sources can be difficult if labelling has been performed using different guidelines. In fact, medical images have many parameters to be considered, such as the resolution of pixels and the size of the Field Of View (FOV). These parameters can be studied in order to

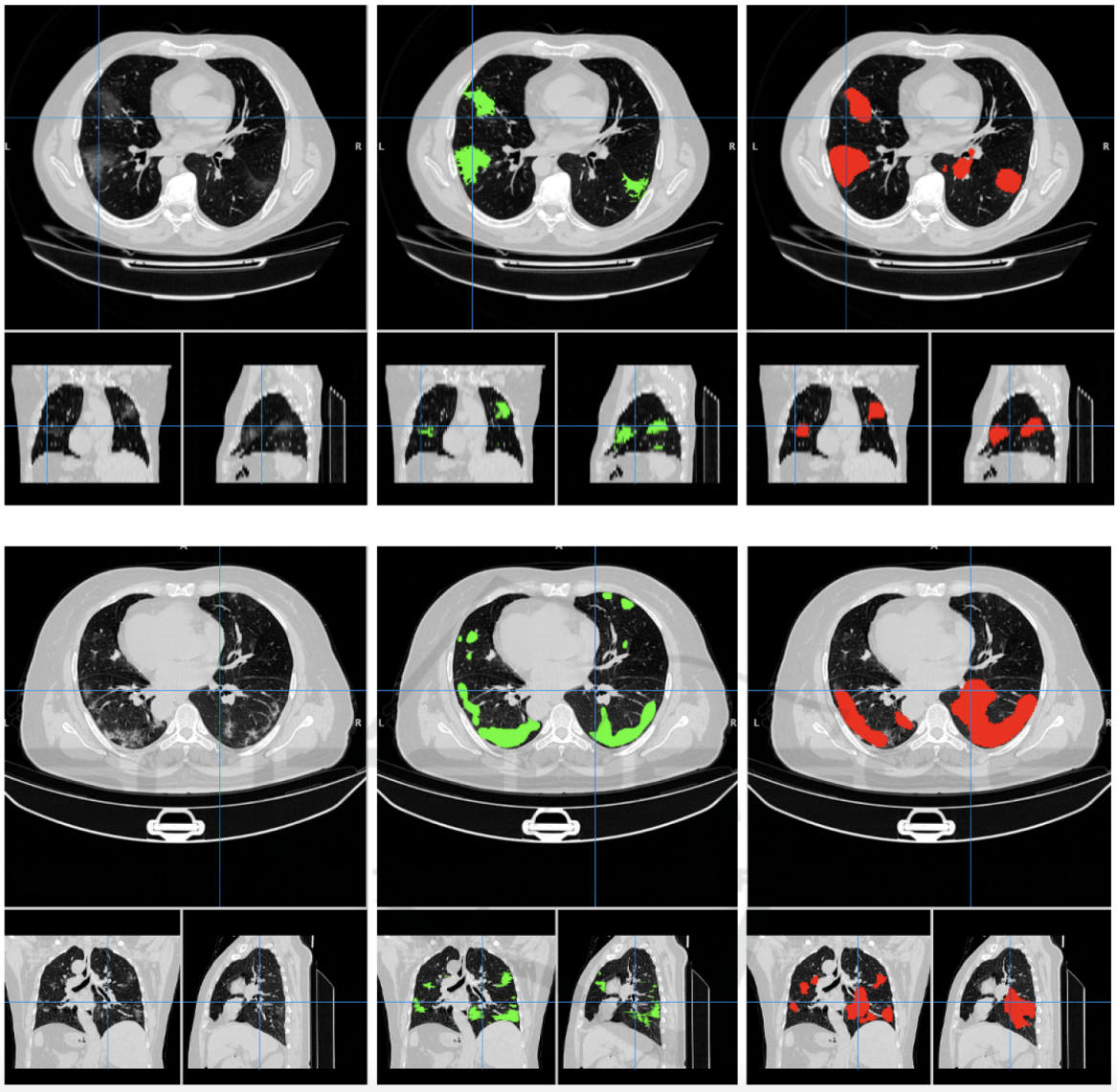


Figure 5.2: Visual comparison between the reference $COVID-19$ lesion masks (green) and the ones predicted (red) by the trained $U-net$ for a representative CT scan of the *MosMed* (first row, $study-0255.nii$) and of the $COVID-19-CT-Seg$ (second row, $coronacases-001.nii$) datasets. The original CT scans are shown on the left as a reference.

attempt a standardization of images from different datasets, by contrast, different annotation styles can not be easily standardized. Since CT image characteristics can be variable, deep learning is a useful method to analyse them and their aggregation. Moreover, $U-nets$ allows a quantification of the volumes of both $COVID-19$ lesions and lungs. On the other hand, the use of deep learning based methods requires a huge amount of homogeneous or harmonized data both to carry out an optimal training process and to implement a fair representation of the population to

be studied.

This preliminary study has been useful to understand which parameters should be considered as the most critical ones in training a neural network model. Lesion labeling and data selection criteria are crucial for this kind of segmentation problems because of the lack of largely populated public datasets, impacting in a relevant way on the performances.

In conclusion, we reviewed all the public available datasets (at the best of our knowledge in April 2021), i.e. *COVID-19 Lung CT Lesion Segmentation Challenge 2020*, *MosMed*, $COVID-19-CT-Seg$ dataset and $TCIA-COVID-19-AR$. We used the Challenge data to train and evaluate a $U-net$ for $COVID-19$ lung lesion segmentation, and we carried out an independent test of the *MosMed* and the $COVID-19-CT-Seg$ datasets, obtaining good performances, as compared to other results available in literature (Ma et al., 2020). We are going to improve our system by adding a module for lung segmentation which could help in quantifying the percentage of lung tissue affected by $COVID-19$ lesions. We also plan to let radiologists evaluate the application of this algorithm on a part of public CT datasets without labelling. Furthermore, segmentation of $COVID-19$ lesions is a starting point for an accurate radiomic analysis for the prediction, based on radiological signs, of the clinical outcome of patients affected by $COVID-19$ pneumonia.

Acknowledgments

This work has been carried out within the Artificial Intelligence in Medicine (AIM) project funded by $INFN$ (CSN5, 2019-2021), `https://www.pi.infn.it/aim`. We are grateful to the staff of the Data Center of the $INFN$ Division of Pisa. We thank the CINECA Italian computing center for making available part of the computing resources used in this paper; in particular, Dr. Tommaso Boccali ($INFN$, Pisa) as PI of PRACE Project Access 2018194658 and a 2021 ISCRA-C grant. Moreover, we thank the EOS cluster of Department of Mathematics "F. Casorati" (Pavia) for computing resources.

Chapter 6

U-nets Cascade

Quantification of pulmonary involvement in COVID-19 pneumonia by means of a cascade of two U-nets: training and assessment on multiple datasets using different annotation criteria

Francesca Lizzi [1,2], Abramo Agosti [6], Francesca Brero [4,5], Raffaella Fiamma Cabini [4,6], Maria Evelina Fantacci [2,3], Silvia Figini [4,11], Alessandro Lascialfari [4,5], Francesco Laruina [1,2], Piernicola Oliva [8,9], Stefano Piffer [7,10], Ian Postuma [4], Lisa Rinaldi [4,5], Cinzia Talamonti [7,10], Alessandra Retico [2]

[1] Scuola Normale Superiore, Pisa, Italy
[2] National Institute of Nuclear Physics (INFN), Pisa Division, Pisa, Italy
[3] Department of Physics, University of Pisa, Pisa, Italy
[4] National Institute of Nuclear Physics (INFN), Pavia Division, Pavia, Italy
[5] Department of Physics, University of Pavia, Pavia, Italy
[6] Department of Mathematics, University of Pavia, Pavia, Italy
[7] National Institute of Nuclear Physics (INFN), Florence Division, Florence, Italy
[8] Department of Chemistry and Pharmacy, University of Sassari, Sassari, Italy
[9] National Institute of Nuclear Physics (INFN), Cagliari Division, Cagliari, Italy
[10] University of Florence, Florence, Italy
[11] Department of Social and Political Science, University of Pavia, Pavia, Italy

I contributed to design the $U-nets'$ architecture and the post-processing of the $U-net_1$ output before to feed it to $U-net_2$. I also suggested the appropriate windowing range of Hounsfield Units. I assisted in evaluating the results and developing a fruitful discussion.

Abstract

Purpose: This study aims at exploiting Artificial Intelligence (AI) for the identification, segmentation and quantification of $COVID-19$ pulmonary lesions. The limited data availability and the annotation quality are relevant factors in training AI-methods. We investigated the effects of using multiple datasets, heterogeneously populated and annotated according to different criteria.

Methods: We developed an automated analysis pipeline, the *LungQuant* system, based on a cascade of two $U-nets$. The first one ($U-net_1$) is devoted to the identification of the lung parenchyma; the second one ($U-net_2$) acts on a bounding box enclosing the segmented lungs to identify the areas affected by $COVID-19$ lesions. Different public datasets were used to train the $U-nets$ and to evaluate their segmentation performances, which have been quantified in terms of the Dice Similarity Coefficient. The accuracy in predicting the CT-SS of the *LungQuant* system has been also evaluated.

Results: Both the volumetric DSC (vDSC) and the accuracy showed a dependency on the annotation quality of the released data samples. On an independent dataset ($COVID-19-CT-Seg$), both the vDSC and the surface DSC (sDSC) were measured between the masks predicted by *LungQuant* system and the reference ones. The vDSC (sDSC) values of 0.95 ± 0.01 and 0.66 ± 0.13 (0.95 ± 0.02 and 0.76 ± 0.18, with $5mm$ tolerance) were obtained for the segmentation of lungs and $COVID-19$ lesions, respectively. The system achieved an accuracy of 90% in CT-SS identification on this benchmark dataset.

Conclusion: We analysed the impact of using data samples with different annotation criteria in training an AI-based quantification system for pulmonary involvement in $COVID-19$ pneumonia. In terms of vDSC measures, the $U-net$ segmentation strongly depends on the quality of the lesion annotations. Nevertheless, the CT-SS can be accurately predicted on independent test sets, demonstrating the satisfactory generalization ability of the *LungQuant*.

Keywords: $COVID-19$, *Chest Computed Tomography, Ground-glass opacities, Segmentation, Machine Learning,* $U-net$

6.1 Introduction

The task of segmenting the abnormalities of the lung parenchyma related to $COVID-19$ infection is a typical segmentation problem that can be addressed with methods based on Deep Learning (DL). CT findings of patients with $COVID-19$ infection may include bilateral distribution of GGO, consolidations, crazy paving patterns, reversed halo sign and vascular enlargement (Carotti et al., 2020). Due to the extremely heterogeneous appearance of $COVID-19$ lesions in density, textural pattern, global shape and location in the lung, an analytical approach is definitely hard to code. The potential of DL-based segmentation approaches is particularly suited in this case, provided that a sufficient number of annotated examples are available for training the models. Few fully automated software tools devoted to this task have been recently proposed (Fang et al., 2021; Lessmann et al., 2021; Ma et al., 2021a). Lessmann et al. (2021) developed a $U-net$ model for lesion segmentation trained on semi-automatically annotated $COVID-19$ cases. The output of this system was then combined with the lung lobe segmentation algorithm reported in Xie et al. (2020). The approach proposed in Fang et al. (2021) implements the automated lung segmentation method provided in the work of Hofmanninger et al. (2020), together with a lesion segmentation strategy based on multiscale feature extraction (Fortin et al., 2018). The specific problem related to the development of fully automated DL-based segmentation strategies with limited annotated data samples has been explicitly tackled by Ma et al. (2021a). The authors studied how to train and evaluate a DL-based system for lung and $COVID-19$ lesion segmentation on poorly populated samples of CT scans. They also made the data publicly available, allowing for a fair comparison with their system. In this work, we present a DL-based fully automated system to segment both lungs and lesions associated with $COVID-19$ pneumonia, the *LungQuant* system, which provides the part of lung volume compromised by the infection. We extended the study proposed by Ma et al. (2021a) focusing our efforts in investigating and discussing the impact of using different datasets and different labelling styles. Data can be highly variable in terms of acquisition protocols and machines when they are gathered from different sources.This poses a serious problem of dependence of the segmentation performances on the training sample characteristics. Despite that advanced data harmonization strategies could mitigate this problem (Fortin et al., 2018), this approach is not applicable in absence of data acquisition information, as it is in this study for the available CT data. Nevertheless, DL methods, when trained with sufficiently large samples of heterogeneous data, can acquire the desired generalization ability by themselves. In our analysis, we implemented an inter-sample cross-validation method to train, test and evaluate

the generalization ability of the *LungQuant* DL-based segmentation pipeline across different available datasets. Finally, we also quantified the effect of using larger datasets to train, validate and test this kind of algorithm.

6.2 Materials and Methods

Datasets

We used only publicly available datasets in order to make our results easily verifiable and reproducible. Five different datasets have been used to train and evaluate our segmentation pipeline. Most of them include image annotations, but each annotation has been associated with patients using different criteria. In Table 6.1, a summary of available labels for each dataset is reported.

The lung segmentation problem has been tackled using a wide representation of the population and three different datasets: the Plethora, the Lung CT Segmentation Challenge and a subset of the MosMed dataset. On the other hand, the number of samples that are publicly available for $COVID-19$ infection segmentation may not be sufficient to obtain good performances on this task. The currently available data, provided along with infection annotations, have been labelled following different guidelines and released in NIfTI format. They do not contain complete acquisition and population information, and they have been stored according to different criteria (see the Supplementary Materials 6.4 for further details). Some of the choices made during the DICOM to NIfTI conversion may strongly affect the quality of data. For example, the MosMed dataset as described by Morozov et al. (2020b) preserves only one slice out of ten during this conversion. This operation results in a significantly loss of resolution with respect to the $COVID-19$ Challenge dataset. Questioning how much such conversion influences the quantitative analysis is important to improve not only the performance but also the possibility of comparing DL algorithm in a fair modality.

LungQuant: a DL based quantification analysis pipeline

The analysis pipeline, which is hereafter referred to as the *LungQuant* system, provides in output the lung and $COVID-19$ infection segmentation masks, the percentage P of lung volume affected by $COVID-19$ lesions and the corresponding CT-SS (CT-SS $=1$ for $P<5\%$, CT-SS $=2$ for $5\% \leq P<25\%$, CT-SS $=3$ for $25\% \leq P<50\%$, CT-SS $=4$ for $50\% \leq P<75\%$, CT-SS $=5$ for $P \geq 75\%$).

Table 6.1: A summary of the datasets used in this study. The CT-Severity Score (CT-SS) information is not available for all datasets, but it can be computed for data which has both lung masks and Ground-Glass Opacifications (GGO) masks.

Dataset name	Lung mask	GGO mask	CT-SS	N. of cases
Plethora Kiser et al. (2020)	Yes	No	No	402
Lung CT Segmentation Challenge Yang et al. (2017)	Yes	No	No	60
COVID-19 Challenge An et al. (2020)	No	Yes	No	199
MosMed Morozov et al. (2020b)	No	No	No	1110
MosMed (annotated subsample)	No	Yes	Inferable	50
MosMed (in-house annotated subsample)	Yes	No	No	91
COVID-19-CT-Seg Ma et al. (2021a)	Yes	Yes	Inferable	10

A summary of our image analysis pipeline is reported in Figure 6.1. The central analysis module is a $U-net$ for image segmentation (Ronneberger et al., 2015) (see Section 6.2), which is implemented in a cascade of two different $U-nets$: the first network, $U-net_1$, is trained to segment the lung and the second one, $U-net_2$, is trained to segment the $COVID$ lesions in the CT scans.

U-net

For both lung and $COVID-19$ lesion segmentation, we implemented a $U-net$ using Keras (2015), a Python DL API that uses Tensorflow as backend. In Figure 6.2, a simplified scheme of our $U-net$ is reported. Each block of layers in the compression path (left) is made by 3 convolutional layers, ReLu activation functions and instance normalization layers. The input of each block is added to the block output in order to implement a residual connection. In the decompression path (right), one convolutional layer has been replaced by a de-convolutional layer to upsample the images to the input size. In the last layer of the $U-nets$, a softmax is applied to the final feature map, and then, the loss is computed.

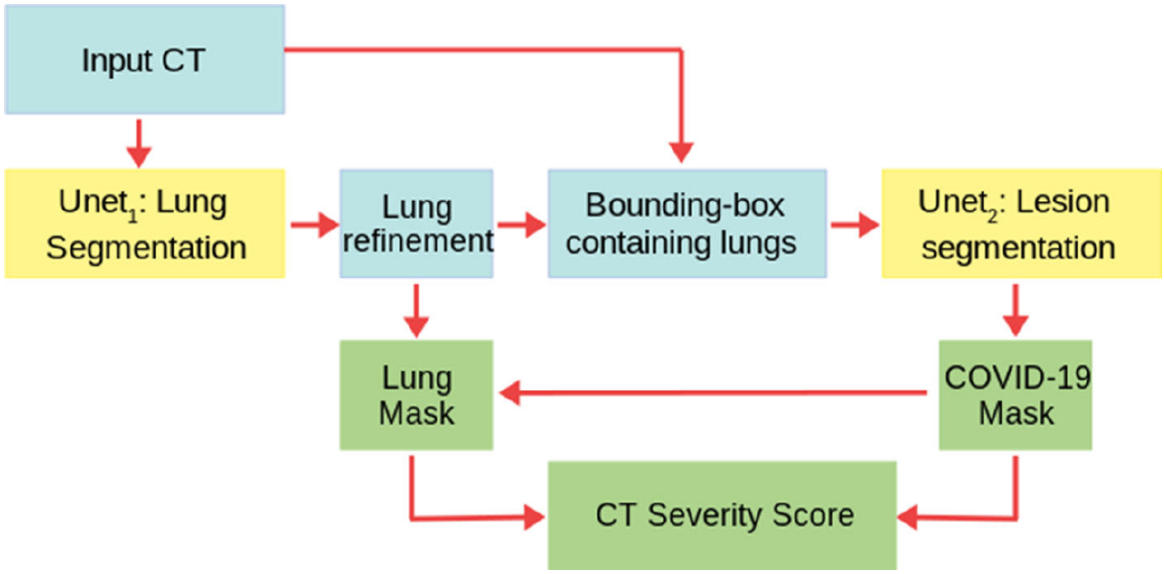


Figure 6.1: A summary of the whole analysis pipeline: the input CT scans are used to train $U-net_1$, which is devoted to lung segmentation; its output is refined by a morphology-based method. A bounding box containing the segmented lungs is made and applied to all CT scans for training $U-net_2$, which is devoted to $COVID-19$ lesion segmentation. Finally, the output of $U-net_2$ is the definitive $COVID-19$ lesion mask, whereas the definitive lung mask is obtained as the union between the outputs of $U-net_1$ and $U-net_2$. The ratio between the $COVID-19$ lesion mask and the lung mask provides the CT-SS for each patient.

The U-net cascade for lesion quantification and severity score assignment

The input CT scans, whose number of slices is highly variable, have been resampled to matrices of $200 \times 150 \times 100$ voxels and then used to train $U-net_1$, which is devoted to lung segmentation, using the three datasets containing original CT scans and lung masks (see Table 6.1). The output of $U-net_1$ was refined using a connected component labelling strategy to remove small regions of the segmented mask not connected with the main objects identified as the lungs. We identified the connected components in the lung masks generated by $U-net_1$, and we excluded those components whose number of voxels was below an empirically fixed threshold (see Supplementary Materials 6.4 for further details). We then built for each CT a bounding box enclosing the refined segmented lungs, adding a conservative padding of $2.5cm$. The bounding boxes were used to crop the training images for $U-net_2$, which has the same architecture as $U-net_1$. Training $U-net_2$ to recognize the $COVID-19$ lesions on a conservative bounding box has two main advantages: it allows to restrict the action volume of the $U-net$ to the region where the lung

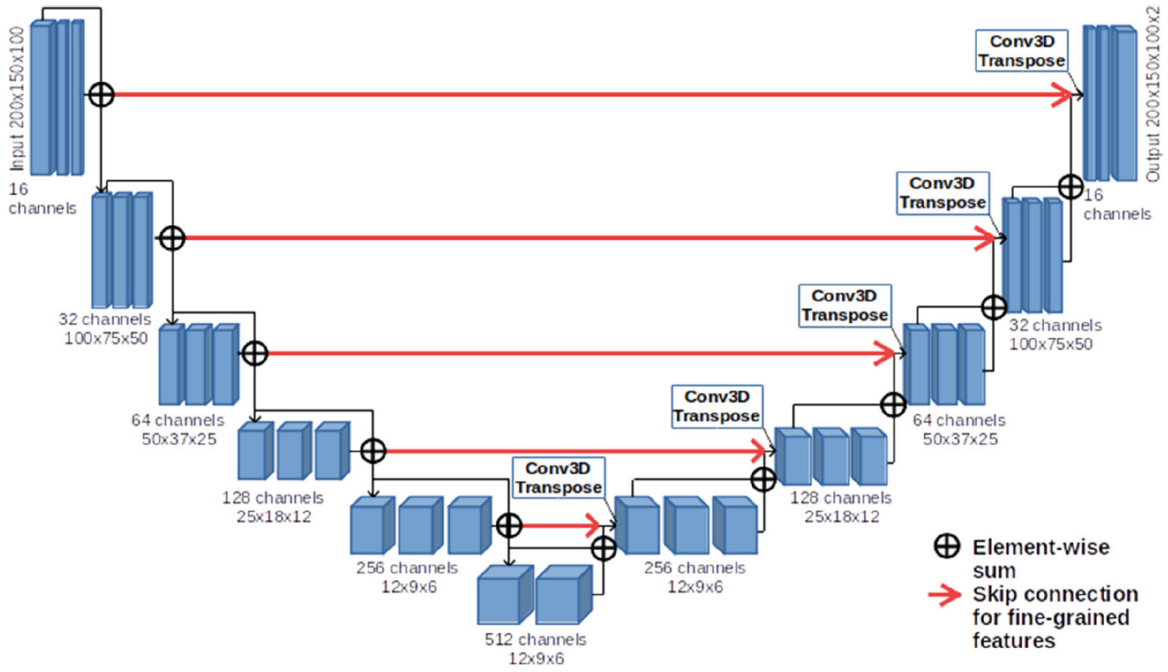


Figure 6.2: $U-net$ scheme: the neural network is made of 6 levels of depth. In the compression path (left), the input is processed through convolutions, activation layers (ReLu) and instance normalization layers, while in the decompression one (right), in addition to those already mentioned, $3D$ Transpose Convolution (de-convolution) layers are also introduced.

parenchyma is supposed to be, thus avoiding false-positive findings outside the chest; it facilitates the $U-net$ training phase, as the dimensions of the lungs of different patients are standardized to focus the $U-net$ learning process on the textural patterns characterizing the $COVID-19$ lesions. The cropped images were resized to a matrix of $200 \times 150 \times 100$ voxels. We applied a windowing on the grey-level values of the CT scans to optimize the image contrast for the two segmentation problems: the $[-1000, 1000]$ HU window range for the $U-net_1$ and the $[-1000, 300]$ HU range for $U-net_2$. The first window highlights the contrast between the lung parenchyma and the surrounding tissues, whereas the second one enhances the heterogeneous structure of the lung abnormalities related to the $COVID-19$ infection. We implemented a data augmentation strategy, relying on the most commonly used data augmentation techniques for DL (see Supplementary Materials 6.4 for further details) to overcome the problem of having a limited amount of labelled data. We transformed the images with rotations, zooming, elastic transformations and adding Gaussian noise.

The $LungQuant$ system returns the infection mask as the output of $U-net_2$ and the lung mask as the union between the output of $U-net_1$ and $U-net_2$. This choice has been made a priori by design, as $U-net_1$ has been trained to segment the lungs

relying on the available annotated data, which are almost totally of patients not affected by $COVID-19$ pneumonia. Thus, $U-net_1$ is expected to be unable to accurately segment the areas affected by GGO or consolidations; as also these areas are part of the lungs, they should be instead included in the mask.

Lastly, once lung and lesion masks have been identified, the *LungQuant* system computes the percentage of lung volume affected by $COVID-19$ lesions as the ratio between the volume of the infection mask and the volume of the lung mask and converts it into the corresponding CT severity score.

Training details and evaluation strategy for the U-nets

Both $U-net_1$ and $U-net_2$ have been evaluated using the volumetric Dice Similarity Coefficient (vDSC). $U-net_1$ has been trained with the vDSC as loss function, while $U-net_2$ has been trained using the sum of the vDSC and a weighted crossentropy as error function in order to balance the number of voxels representing lesions and the background (see Supplementary Materials 6.4 for further details). The performances of the whole system have been evaluated also with the surface Dice Similarity Coefficient (sDSC) for different values of tolerance (Kiser et al., 2021).

Cross-validation strategy

To train, validate and test the performances of the two $U-nets$, we partitioned the datasets into training, validation and test sets. We then evaluated the network performance separately and globally. $U-net_2$ has been trained twice, i.e. on the 60% and 90% of the CT scans of $COVID-19$ Challenge and Mosmed datasets to investigate the effect of maximizing the training set size on the lesion segmentation. The amount of CT scan used for train, validation and test sets for each $U-net$ is reported in Table 6.2. To evaluate the ability of the trained networks to predict the percentage of the affected lung parenchyma and thus the CT-SS classification, we used a completely independent set consisting of 10 CT scans from the COVID-19-CT-Seg dataset, which is the only publicly available dataset containing both lung and infection mask annotations.

6.3 Results

In this section, we report, first, the performance achieved by $U-net_1$ and $U-net_2$, then, the quantification performance of the integrated *LungQuant* system, evaluated on a completely independent test set. We trained both the $U-nets$ for 300 epochs

Table 6.2: Number of CT scans assigned to the train, validation (val) and test sets used during the training and performance assessment of the $U-net_1$ and the $U-net_2$ networks.

U-net$_1$	Train	Validation	Test
Plethora	319	40	40
MosMed (91 CT-0)	55	18	18
LCTSC	36	12	12
COVID-19-CT-Seg	-	-	10
U-net$_2^{60\%}$	Train (60%)	Val (20%)	Test
COVID-19 Challenge	119	40	40
MosMed (50 CT-1)	30	10	10
COVID-19-CT-Seg	-	-	10
U-net$_2^{90\%}$	Train (90%)	Val (10%)	Test
COVID-19 Challenge	179	20	-
MosMed (50 CT-1)	45	5	-
COVID-19-CT-Seg	-	-	10

on a NVIDIA V100 GPU using ADAM as optimizer and we kept the models trained at the epoch where the best evaluation metric on the validation set was obtained.

U-net$_1$: Lung segmentation performance

$U-net_1$ for lung segmentation was trained and validated using three different datasets, as specified in Table 6.2. Then, we tested $U-net_1$ on each of the three independent test sets and we reported in Table 6.3 the performance achieved in terms of vDSC, computed between the segmented masks and the reference ones, both separately for each dataset and globally.

The evaluation of the lung segmentation performances was made in three cases: (1) on CT scans and masks resized to the $200 \times 150 \times 100$ voxel array size; (2) on CT scans and masks in the original size before undergoing the morphological refinement; (3) on CT scans and masks in the original size and after the morphological refinement. Even if segmentation refinement has a small effect on vDSC, since it is a volume-based metric, as shown in Table 6.3, it is a fundamental step to allow the definition of precise bounding boxes enclosing the lungs and thus to facilitate the $U-net_2$

Table 6.3: Performances achieved by $U-net_1$ in lung segmentation on different test sets, evaluated in terms of the vDSC at three successive stages of the segmentation procedure.

Test set	Masks of U-net size (vDSC)	Masks before refinement (vDSC)	Masks after refinement (vDSC)
Plethora	0.96 ± 0.02	0.95 ± 0.02	0.95 ± 0.04
MosMed	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02
LCTSC	0.96 ± 0.03	0.95 ± 0.03	0.96 ± 0.01
COVID-19-CT-Seg	0.96 ± 0.01	0.95 ± 0.01	0.95 ± 0.01

learning process.

U-net$_2$: COVID-19 lesion segmentation performance

$U-net_2$ for $COVID-19$ lesion segmentation has been trained and evaluated separately on the COVID-19-Challenge dataset and on the annotated subset of the MosMed dataset, following the train/validation/test partitioning reported in Table 6.2. The segmentation performances achieved on the test sets are reported in terms of the vDSC in Table 6.4, according to the cross-sample validation scheme.

Table 6.4: Performances achieved by $U-net_2$ in $COVID-19$ lesion segmentation, evaluated in terms of the vDSC.

U-net	Trained on	Test set	Test set U-net size (vDSC)	Original CT size (vDSC)
U-net$_2^{60\%}$	COVID-19 Challenge	COVID-19 Challenge	0.51 ± 0.24	0.51 ± 0.25
	COVID-19 Challenge	MosMed	0.39 ± 0.19	0.40 ± 0.19
	MosMed	MosMed	0.54 ± 0.22	0.55 ± 0.22
	MosMed	COVID-19 Challenge	0.25 ± 0.23	0.25 ± 0.23
	COVID-19 Challenge + MosMed	COVID-19 Challenge + MosMed	0.49 ± 0.21	0.50 ± 0.21
U-net$_2^{90\%}$	COVID-19 Challenge + MosMed	COVID-19 Challenge + MosMed	0.64 ± 0.23	0.65 ± 0.23

The composition of the train and test sets is reported in Table 6.2.

As expected, the $U-net_2$ performances are higher when both the training set and independent test sets belong to the same data cohort. By contrast, when a $U-net_2$ is trained on COVID-19-Challenge data and tested on MosMed (and the

other way around), performances significantly decrease. This effect is related to different criteria used to both collect and annotate the data. We obtained a better result with the $U-net2$ trained on the COVID-19 Challenge dataset and tested on the MosMed test set, since the network has been trained on a larger data sample and hence it has a higher generalization capability. The best segmentation performances have been obtained by the $U-net2$ trained using the 90% of the available data, $U-net_2^{90\%}$, which reaches a vDSC of 0.65 ± 0.23 on the test set. This result suggests the need to train $U-net$ models on the largest possible data samples in order to achieve higher segmentation performance.

Evaluation of the quantification performance of the *LungQuant* system on a completely independent set

Evaluation of lung and COVID-19 lesion segmentations

Once the two $U-nets$ have been trained and the whole analysis pipeline has been integrated into the *LungQuant* system, we tested it on a completely independent set (COVID-19-CT-Seg dataset) of CT scans. The performances of the whole process were quantified both in terms of vDSC and sDSC with tolerance values of 1, 5 and $10mm$ (Table 6.5). A very good overlap between the predicted and reference lung masks is observable in terms of vDSC, whereas the sDSC values are highly dependent on tolerance values, ranging from moderate to very good agreement measures. Regarding the lesion masks, a moderate overlap is measured between the predicted and reference lesion masks in terms of vDSC, whereas the sDSC returns measures extremely dependent on tolerance values that span from limited to moderately good and ultimately satisfactory performances for tolerance values of $1mm$, $5mm$ and $10mm$, respectively. Figure 6.3 allows for a visual comparison between the lung and lesion masks provided by the *LungQuant* system integrating $U-net_2^{90\%}$ and the reference ones.

Percentage of affected lung volume and CT-SS estimation

The lung and lesion masks provided by the *LungQuant* system can be further processed to derive the physical volumes of each mask and the ratios between the lesion and lung volumes. We show in Figure 6.4 the relationship between the percentage of lung involvement as predicted by the *LungQuant* system vs. the corresponding values for the reference masks of the fully independent test set COVID-19-CT-Seg, for both the *LungQuant* systems with the $U-net_2^{60\%}$ and the $U-net_2^{90\%}$. Despite the limited range of samples to carry out this test, an agreement between the *LungQuant* sys-

Table 6.5: Performances of the *LungQuant* system on the independent COVID-19-CT-Seg test dataset. The vDSC and sDSC computed between the lung and lesion reference masks and those predicted by the *LunQuant* system are reported.

Metrics	vDSC	sDSC (1 mm)	sDSC (5 mm)	sDSC (10 mm)
	Lung segmentation			
LungQuant (U-net260%)	0.96 ± 0.01	0.66 ± 0.09	0.95 ± 0.02	0.98 ± 0.01
LungQuant (U-net290%)	0.95 ± 0.01	0.65 ± 0.09	0.95 ± 0.02	0.98 ± 0.01
	Infection Segmentation			
LungQuant (U-net260%)	0.62 ± 0.09	0.29 ± 0.06	0.75 ± 0.11	0.90 ± 0.09
LungQuant (U-net290%)	0.66 ± 0.13	0.36 ± 0.13	0.76 ± 0.18	0.87 ± 0.13

tem output and the reference values is observed for both $U - net_2^{60\%}$ and $U - net_2^{90\%}$. In terms of the Mean Absolute Error (MAE) among the estimated and the reference percentages of affected lung volume (P), we obtained a Mean Absolute Error equal to $MAE = 4.6\%$ for the *LungQuant* system with $U - net_2^{60\%}$ and $MAE = 4.2\%$ for the system with $U - net_2^{90\%}$.

The accuracy in assigning the correct CT-SS class is reported in Table 6.6, together with the number of misclassified cases, for the 10 cases of the COVID-19-CT-Seg dataset. The best accuracy achieved by LungQuant is of 90% with $U - net_2^{90\%}$. In all cases, the system misclassifies the examples by 1 class at most.

Table 6.6: Classification performances of the whole system in predicting CT-Severity Score on MosMed and COVID-19-CT-Seg datasets. The number of misclassified cases is reported.

U-net	Dataset	Accuracy	Misclassified by 1 class	Misclassified by 2 classes
U-net$_2^{60\%}$	COVID-19-CT-Seg	6/10	4/10	0
U-net$_2^{90\%}$	COVID-19-CT-Seg	9/10	1/10	0

6.4 Discussion and Conclusion

We developed a fully automated quantification pipeline, the *LungQuant* system, for the identification and segmentation of lungs and pulmonary lesions related to $COVID-19$ pneumonia in CT scans. The system returns the $COVID-19$ related lesions, the lung mask and the ratio between their volumes, which is converted into

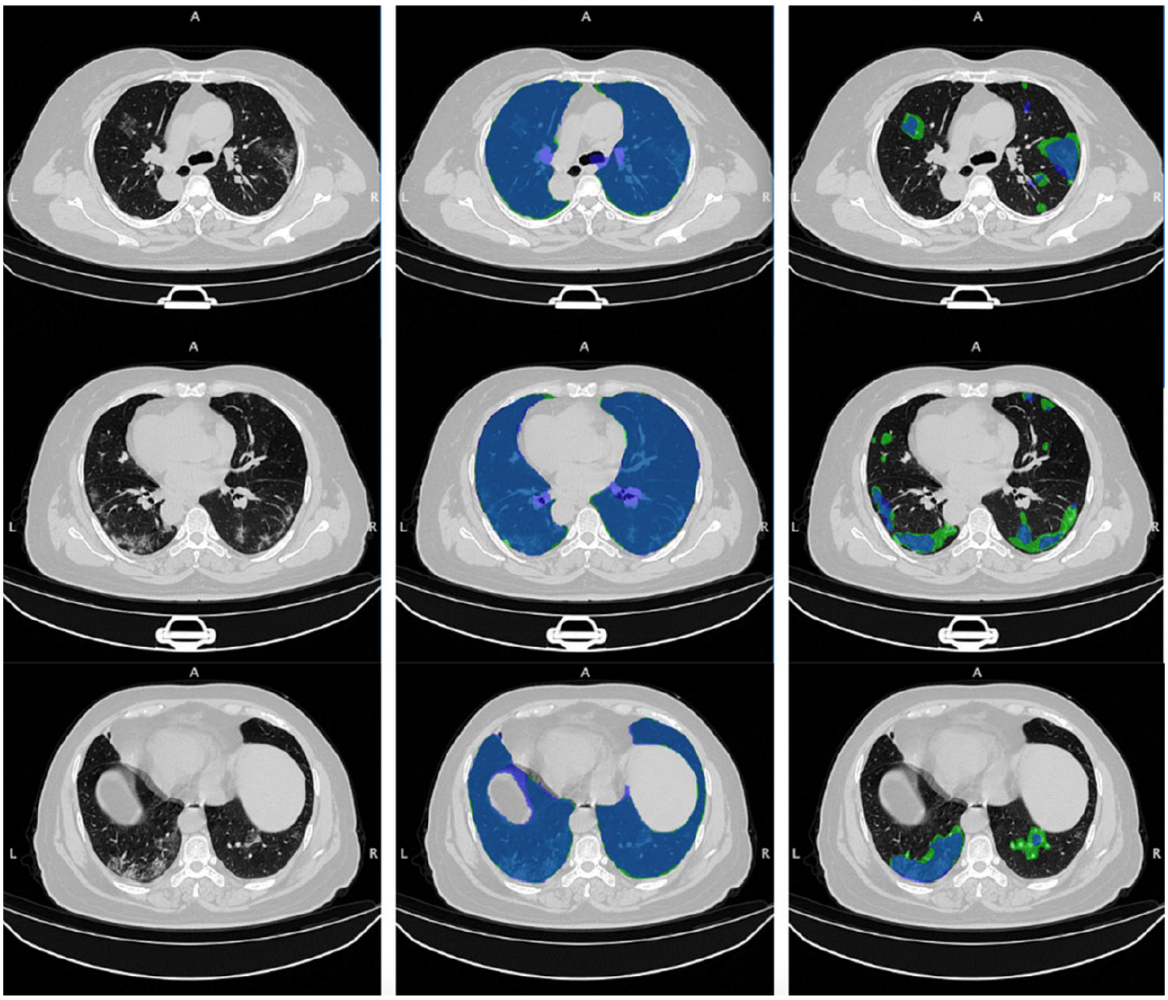


Figure 6.3: On the rows: three axial slices of the first CT scan on the COVID-19-CT-Seg test dataset ($4coronacases001.nii$) are shown. On the columns: original images (left); overlays between the predicted and the reference lung (centre) and $COVID-19$ lesion (right) masks. The reference masks are in green, while the predicted ones, obtained by the *LungQuant* system integrating $U-net_2^{90\%}$, are in blue.

a CT Severity Score. The performance obtained against a voxel-wise segmentation ground truth was evaluated in terms of the vDSC, which provides a measure of the overlap between the predicted and the reference masks. The *LungQuant* system achieved a vDSC of 0.95 ± 0.01 in the lung segmentation task and of 0.66 ± 0.13 in segmenting the $COVID-19$ related lesions on the fully annotated publicly available benchmark COVID-19-CT-Seg dataset of 10 CT scans. The *LungQuant* has been evaluated also in terms of sDSC for different values of tolerance. The results obtained at a tolerance of $5mm$ equal to 0.76 ± 0.18 are satisfactory for our purpose, given the heterogeneity of the labelling process. Regarding the correct assignment of the CT-SS, the *LungQuant* system showed an accuracy of 90% on the completely independent

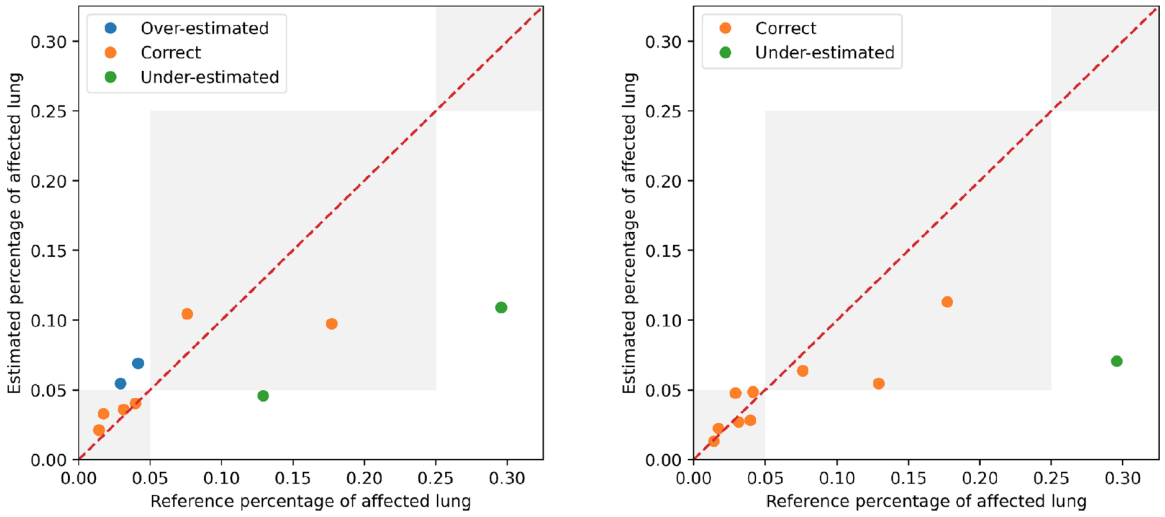


Figure 6.4: Estimated percentages P of affected lung volume versus the ground truth percentages, as obtained by the *LungQuant* system integrating $U - net_2^{60\%}$ (left) and $U - net_2^{90\%}$ (right). The grey areas in the plot backgrounds guide the eye to recognize the CT-SS values assigned to each value of P (from left to right: CT-SS = 1, CT-SS = 2, CT-SS = 3)

test set COVID-19-CT-Seg.

Despite that this result is encouraging, it was obtained on a rather small independent test set; thus, a broader validation on larger data sample with more heterogeneous composition in terms of disease severity is required. Training DL algorithms requires a huge amount of labelled data. The lung segmentation task has been made feasible in this work thanks to the use of lung CT datasets collected for purposes different from the study of $COVID-19$ pneumonia. Training a segmentation system on these samples had the effect that when we use the trained network to process the CT scan of a patient with $COVID-19$ lesions, especially in case ground glass opacities and consolidation are very severe, the lung segmentation is not accurate anymore. In order to overcome this problem, the proposed *LungQuant* system returns a lung mask which is the logical union between the output mask of the $U - net_1$ and the infection mask generated by the $U - net_2$. The *LungQuant* system can actually be improved whether lung masks annotation are available on subjects with $COVID-19$ lesions. A similar problem occurs for the segmentation of ground glass opacities and consolidations. The available data, in fact, are very unbalanced with respect to the severity of $COVID-19$ disease, and hence, the accuracy in segmenting the most severe case is worse. The current lack of a large dataset, collected by paying attention to adequately represent all categories of disease severity, limits the possibility to carry out accurate training of AI-based models. Finally, we found that

the difference in the annotation and collection guidelines among datasets is an issue. Processing aggregated data from different sources can be difficult if labelling has been performed using different guidelines. CT scans should contain the acquisition parameters, usually stored in the DICOM header, when they are published. The lack of this information is a drawback of our study. If we had that data, we could study more in detail the dependence of the *LungQuant* performances on specific acquisition protocols or scanners. On the contrary, even with this information, it would not be possible to standardize the different annotation styles. The results of *LungQuant* (last 2 rows of Table 6.4) demonstrate its robustness across different datasets even without a dedicated preprocessing step to account for this information.

Acknowledgments

This work has been carried out within the Artificial Intelligence in Medicine (AIM) project funded by $INFN$ (CSN5, 2019-2021), `https://www.pi.infn.it/aim`. We are grateful to the staff of the Data Center of the $INFN$ Division of Pisa. We thank the CINECA Italian computing center for making available part of the computing resources used in this paper; in particular, Dr. Tommaso Boccali ($INFN$, Pisa) as PI of PRACE Project Access 2018194658 and a 2021 ISCRA-C grant. Moreover, we thank the EOS cluster of Department of Mathematics "F. Casorati" (Pavia) for computing resources.

Supplementary Materials - Additional descriptions of Materials and Methods

Characteristics of the public datasets used in the study

The Plethora dataset

The *PleThora* dataset (Kiser et al., 2020) is a chest CT scan collection with thoracic volume and pleural effusion segmentations, delineated on 402 CT studies of the Non-Small Cell Lung Cancer (NSCLC) radiomics dataset, available through the The Cancer Imaging Archive (TCIA) repository Clark et al. (2013). This dataset has been made publicly available to facilitate improvement of the automatic segmentation of lung cavities, which is typically the initial step in the development of automated or semi-automated algorithms for chest CT analysis. In fact, automatic lung identification struggles to perform consistently in subjects with lung diseases. The PleThora lung annotations have been produced with a $U-net$ based algorithm trained on chest CT of subjects without cancer, manually corrected by a medical student and revised by a radiation oncologist or a radiologist.

The 2017 Lung CT Segmentation Challenge dataset

The *Lung CT Segmentation Challenge* (LCTSC) dataset consists of CT scans of 60 patients, acquired from 3 different institutions and made publicly available in the context of the *2017 Lung CT Segmentation Challenge* (Yang et al., 2017). Since the aim of this challenge was to foster the development of auto-segmentation methods for organs at risk in radiotherapy, the lung annotations followed the RTOG 1106 contouring atlas.

The 2020 COVID-19 Lung CT Lesion Segmentation Challenge dataset

The *2020 COVID-19 Lung CT Lesion Segmentation Challenge* dataset ($COVID-19$ Challenge) is a public dataset consisting of unenhanced chest CT scans of 199 patients with positive RT-PCR for SARS-CoV-2 (An et al., 2020). Each CT is accompanied with the ground truth annotations for $COVID-19$ lesions. Data has been provided in NIfTI format by The Multi-national NIH Consortium for CT AI in $COVID-19$ via the NCI TCIA public website (Clark et al., 2013). Annotations have been made using a $COVID-19$ lesion segmentation model provided by NVIDIA, which takes a full CT chest volume and produces pixel-wise segmentations of $COVID-19$ lesions. These segmentations have been adjusted manually by a certified radiologists board, in order to give $3D$ consistency to lesion masks. The dataset annotation was

made possible through the joint work of Children's National Hospital, NVIDIA and National Institutes of Health for the *COVID-19-20 Lung CT Lesion Segmentation Grand Challenge*.

The dataset and the annotations have been made available in the context of a MICCAI-endorsed international challenge (`https://covid-segmentation.grand-challenge.org/`) which had the aim to compare AI-based approaches to automated segmentation of $COVID-19$ lung lesions.

The MosMed dataset

MosMed (Morozov et al., 2020b) is a $COVID-19$ chest CT dataset collected by the Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department. It includes CT studies taken from 1110 patients. Each study is represented by one series of images reconstructed into soft tissue mediastinal window. MosMed provides 5 labeled categories, based on the percentage of lung parenchyma affected by $COVID-19$ lesions. The 5 categories of lung involvement and their correspondence to the CT-SS scale are described in Table 6.1. The first category (CT-0) contains cases with normal lung tissue and no CT-signs of viral pneumonia, whereas the other categories contain GGO (CT-1 and CT-2) and both GGO and regions of consolidation in the higher classes (CT-3 and CT-4).

Supplementary Table 6.1: MosMed severity categories defined on the basis of the percentage P of lung volume affected by $COVID-19$ lesions. The correspondence to the CT-SS scale is reported.

MosMed CT category	N. of cases	Percentage P of involved lung parenchyma	Corresponding CT-SS
0	254	$P = 0$	0
1	684	$0 < P \leq 25$	1, 2
2	125	$25 < P \leq 50$	3
3	45	$50 < P \leq 75$	4
4	2	$75 < P \leq 100$	5

A small subset of class CT-1 cases (50 patients) had been annotated by expert radiologists with the support of MedSeg software (2020 Artificial Intelligence AS). The annotations consist of binary masks in which white voxels represent both ground-glass opacifications and consolidations. Both CT scans and annotations were provided in NIfTI format. During the $DICOM-to-NIfTI$ conversion process, only

one slice out of ten was preserved and, as a result, MosMed CT scans have a reduced total number of slices with respect to the other datasets.

The COVID-19-CT-Seg dataset

The *COVID-19-CT-Seg* dataset is a collection of CT scans taken from the Coronacases Initiative and Radiopaedia (Ma et al., 2021a). It contains 20 CT scans tested positive for $COVID-19$ infection. This public dataset contains both lung and infection annotations. The ground truth has been made in three steps: first, junior radiologists ($1-5$ years of experience) delineated lungs and infections annotations, then two radiologists ($5-10$ years of experience) refined the labels and finally the annotations have been verified and optimized by a senior radiologist (more than 10 years of experience in chest radiology). The annotations have been produced with the ITK-SNAP software. Ten CT images of this dataset were provided in 8-bit depth, therefore, we decided to not use them.

Additional training details and evaluation strategy for the U-nets

Evaluation metrics

The segmentation performances for both $U-nets$ have been evaluated with the volumetric Dice Similarity Coefficient (vDSC), computed between the true mask volume (V_{true}) and the predicted mask volume ($V_{predict}$), and with the surface Dice Similarity Coefficient (sDSC), computed between the true surface (S_{true}), and the predicted one defined, ($S_{predict}$) (Kiser et al., 2021), as follows;

$$vDice_{metric} = \frac{2 \cdot |V_{true} \cap V_{predict}|}{|V_{true}| + |V_{predict}|} \tag{2}$$

$$sDice_{metric} = \frac{2 \cdot |S_{true} \cap S_{predict}|}{|S_{true}| + |S_{predict}|} \tag{3}$$

The loss function used to train the $U-net_1$ for lung segmentation is the vDSC loss, defined as follows

$$vDice_{loss} = 1 - \frac{2 \cdot |M_{true} \cap M_{predict}|}{|M_{true}| + |M_{predict}|} \tag{4}$$

and computed only on the foreground (white voxels). We used this strategy in order to avoid giving excessive weight to the background (black voxels), since the number of black and white voxels is quite unbalanced in favor of the former.

For $U - net_2$, we used a loss function (L) consisting of the sum of the vDSC loss and a weighted cross-entropy (CE), defined as follows:

$$L = vDice_{loss} + CE_{weighted} \tag{5}$$

$$CE_{weighted} = w(x) \sum_{x \in \Omega} log(M_{true}(x) \cdot M_{predict}(x)) \tag{6}$$

where $w(x)$ is the weight map which takes into account the frequency of white voxels, x is the current sample and Ω is the training set.

Since the background class is larger than the foreground class on the order 10^3, we computed the weight map $w(x)$ for each ground-truth segmentation to increase the relevance of the underrepresented class, following the approach described in Phan and Yamamoto (2020). The weight map was defined as $w(x) = w_0 / f_j$ where f_j is the average number of voxels of the j^{th} class over the entire training data set $(j = 0, 1)$ and w_0 is the the average between the frequencies f_j.

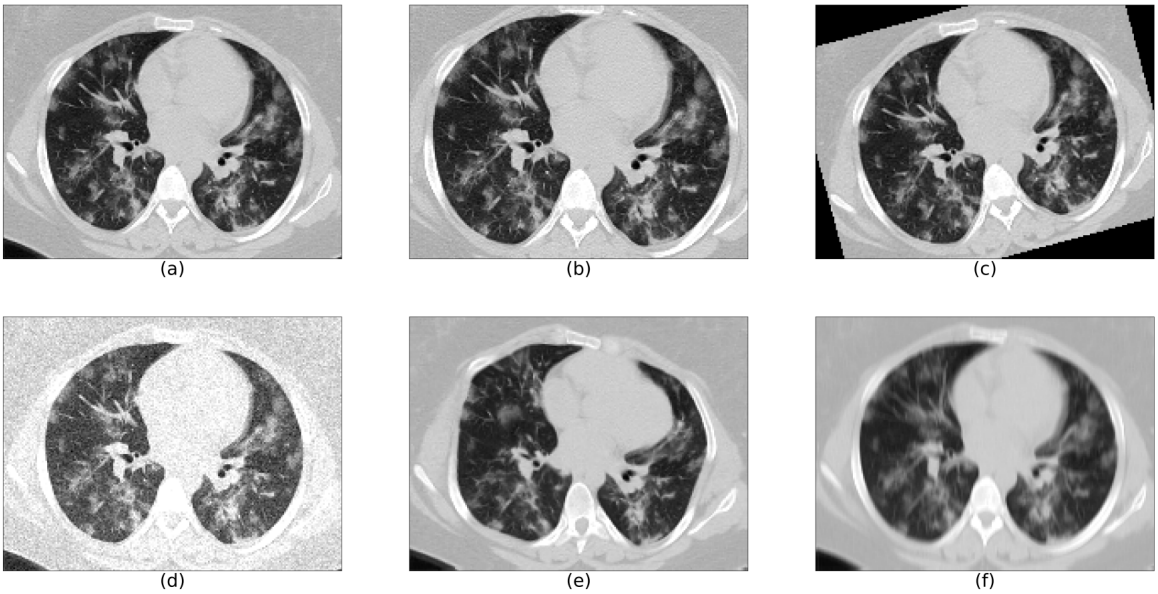
Data augmentation

Data augmentation is a strategy to increase the size of the training set by synthetically generating additional training images through geometric transformations. This technique is particularly important to improve the generalization capability of the model, especially in the case of a limited number of training samples. In our work, we applied data augmentation during the data pre-processing phase (after defining the bounding boxes enclosing the segmented lungs) in order to generate a fixed number of augmented images for each original data. We chose an augmentation factor equal to 2 which means that the number of artificially generated images is twice the number of the original training set. For each image in the training set, two of the following geometric transformations were randomly chosen:

- Zooming. The CT image and the ground truth masks were zoomed in the axial plane, using a third-order spline interpolation and the *k*-nearest neighbor method, respectively. The zooming factor was randomly chosen among the following values: 1.05, 1.1, 1.15, 1.2.
- Rotation. The CT image and the ground truth mask were rotated in the axial plane, using a third-order spline interpolation and the k-nearest neighbor method, respectively. The rotation angle was randomly sampled among the following values: $-15°$, $-10°$, $-5°$, $5°$, $10°$, $15°$.

- Gaussian noise. An array of noise terms randomly drawn from a normal distribution was added to the original CT image. For each image, the mean of the Gaussian distribution was randomly sampled in the $[-400, 200]$ HU range and the standard deviation randomly chosen among 3 values: 25, 50, 75 HU.
- Elastic deformation. An elastic distortion was applied to the original $3D$ CT and mask arrays following the approach of Simard et al. (2003). This transformation has two parameters: the elasticity coefficient which we fixed to 12 and the scaling factor, fixed to 1000.
- Motion blurring. Slice by slice, we convolved the CT image with a linear kernel (i.e. ones along the central row and zero elsewhere for a matrix of size $k \times k$) through the function $filter2D$, defined in the OpenCV Python library (Bradski, 2000), keeping the output image size the same as the input image. The filter is applied with a kernel size of 4, 3, and 3, in the anterior-posterior, latero-lateral and cranio-caudal direction, respectively.

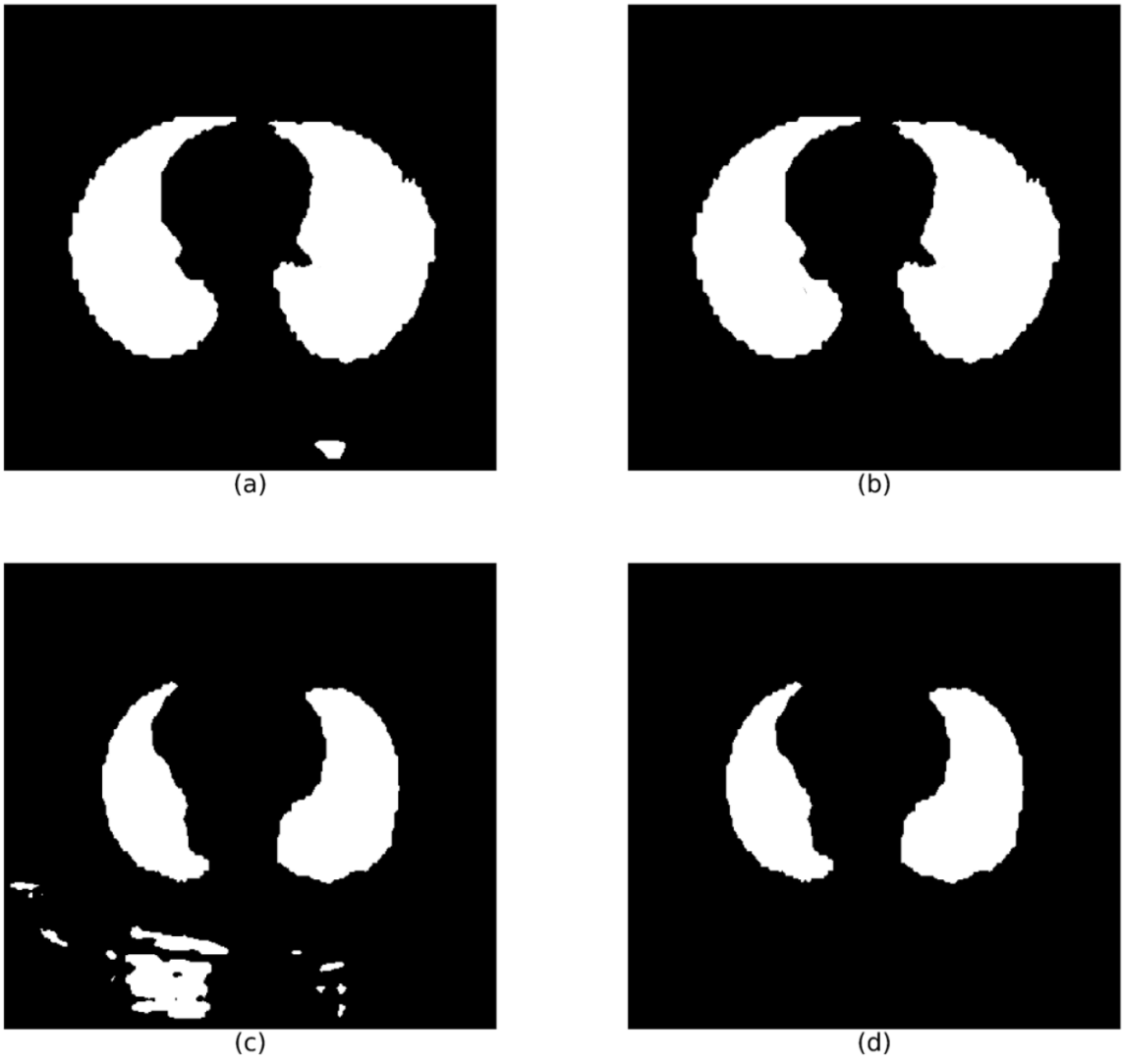
An example of the application of these augmentation techniques to one CT scan of the dataset is provided in Figure 6.1.



Supplementary Figure 6.1: Data augmentation to increase the diversity of dataset: a) Image without data augmentation; b) Zooming; c) Rotation; d) Gaussian noise; e) Elastic deformations; f) Motion blurring.

Morphological refinement of U-net$_1$ lung segmentation

In order to remove false-positive regions (i.e. voxels misclassified as lung parts), at first, we identified the connected components in the lung masks generated by $U - net_1$, then, we excluded those components whose number of voxels was below an empirically-fixed threshold. This threshold was set to the 40% of the foreground mask, and it was reduced to 30% whether the resulting number of voxels was found to be lower than the 65% of the initial mask provided by $U - net_1$. Figure 6.2 shows some examples of how this procedure works on real CT scans.



Supplementary Figure 6.2: Morphological refinement of the $U - net_1$ output: a) and c) lung masks as generated by $U - net_1$; b) and d) refined masks after the connected component selection.

Generation of a set of reference lung segmentation for model training

As reported in Table 6.1, the available datasets with lung mask annotations, which were necessary to train the $U-net$ for lung segmentation, are mainly of subjects affected by lung cancer (Plethora and LCTSC datasets). To complement this sample with subjects without lesions, and, at the same time, to expose the $U-net$ to the acquisition characteristics of the MosMed CT scans, we generated the lung mask annotations for a subset of subjects of the CT-0 MosMed category, i.e. that of subjects without $COVID-19$ lesions.

An in-house lung segmentation algorithm was developed for this purpose and implemented in Matlab (The MathWorks, Inc.). It is based on the following steps: 1) CT windowing in the $[-1000, 1000]$ HU range; 2) rough segmentation of the lungs on a central coronal slice (Otsu binary thresholding and removal of components connected with the image border) to define the minimum and maximum axial coordinates of the lung region; 3) $2D$ rough segmentation of the lungs on each axial slice (same procedure as the previous step) to generate a $3D$ seed mask for the following step; 4) segmentation of the lung parenchyma by an active contour model (*activecontour* Matlab function); 5) filling holes (e.g. vessels and airway walls) with $3D$ morphological operators (*imclose* Matlab function).

This algorithm, which accurately segments the lung parenchyma in absence of lesions, has very limited performance on CT scans of subjects with $COVID-19$ lesions.

Chapter 7

Transfer Learning

Development and validation of deep learning soft-tissue-sarcoma distant metastasis prediction based on transfer learning and fine-tuning with MR, CT and dose images

Stefano Piffer [1,2], Mauro Loi [3], Daniela Greto [3], Cinzia Talamonti [1,2]

[1] National Institute of Nuclear Physics (INFN), Florence Division, Florence, Italy
[2] University of Florence, Florence, Italy
[3] Radiation Oncology Unit, Careggi University Hospital, Florence, Italy

I contributed to think and design the experiment. I personally created and curated the entire database from scratch and collaborated with the expert radiotherapist (100%). I performed data preprocessing (images co-registration, resolution adjustment, tumor contours propagation) and I designed, implemented, and trained 100% of the deep learning models. I also drafted the article and prepared 100% of the figures.

Abstract

Background: Soft-Tissue-Sarcomas (STSs) are uncommon, heterogeneous malignant tumors and their clinical management is particularly challenging. Accurate and precise STS patients' stratification play an important role in clinical diagnosis and decision making for patient treatment. Recent development on deep learning has shown great progress also in medical fields but the main limitation remains the small labeled dataset for training. To overcome this drawback transfer learning and fine-tuning have been investigated. The goal of this study is to predict STS patients' outcome to radiotherapy, in terms of distant metastasis development.

Methods: In this single-center analysis, 61 STS patients between 2011 and 2020 were retrospectively enrolled. We designed a pipeline employing transfer learning and fine-tuning a pre-trained $VGG-19$ network. The prediction model was trained, validated and tested on ten different combination of the available multimodal images (CT, dose distribution, $T1$-weighted, $T2$-weighted and contrast-enhanced $T1$-weighted MRI) within ten-fold cross-validation. Accuracy, sensitivity specificity, precision and $F1$-score for slice-based prediction were assessed.

Results: The best performance was achieved considering Dose – $T2$-weighted – contrast-enhanced $T1$-weighted multimodal images combination. In this configuration, the averaged slice-based prediction accuracy was 0.93 ± 0.02, 0.92 ± 0.02, and 0.90 ± 0.03 for training, validation, and test respectively, under ten-fold cross-validation.

Conclusion: This study demonstrated the opportunity to use deep learning coupled with transfer learning and fine-tuning to predict distant metastasis from an eccentric combination of multimodal images. Despite the result is still being preliminary, it demonstrated the feasibility and the efficacy of the proposed workflow on predicting post-RT patient outcome. This could potentially improve the clinical management of STS patients.

Keywords: *soft-tissue-sarcoma, transfer learning, fine-tuning, distant metastasis, magnetic resonance imaging, dose distribution*

7.1 Introduction

Soft-Tissue-Sarcomas (STSs) represent a rare and heterogeneous group of tumors, with more than 100 histological subtypes and account for 1% of solid cancers in adults (Igrec and Fuchsjäger, 2021; Coindre et al., 2001; Bray et al., 2018). Even their occurrence is heterogeneous, but the limbs are the most common primary site for a Soft-Tissue-Sarcoma (Gao et al., 2021).

Because of the diversity in presentation and outcome within the spectrum of STS, several prognostic instruments have been developed to classify patients with these tumors into risk groups to optimize their management. Traditional pathology approaches and molecular genetic assays have played a crucial role in the classification of STS. An accurate histological diagnosis and an assessment of the risk of relapse are critical for delineating treatment strategies. Pretreatment pathologic assessment consists of percutaneous core needle biopsy and limits the classification of the entire tumor with few tissue samples. A complete and in-depth histological analysis takes place only after surgery which is an advanced step in the therapeutic process and may differ from the preliminary grade provided on biopsy specimens (Crombé et al., 2022; Corino et al., 2018; Schneider et al., 2017).

Interdisciplinary management of extremity and truncal soft tissue sarcoma includes a multimodal combination of treatments, such as margin-negative surgical resection, external beam-radiation therapy and systemic chemotherapy. Generally accepted guidelines suggest applying pre-operative external beam Radiotherapy (RT), conventionally fractionated in $25-28$ fractions of $1.8-2Gy$ to a total dose of $50-50.4Gy$ in $5-6$ weeks. Post-operative RT instead provides $60-66Gy$ delivered in $1.8-2Gy$ fractions over 6 to 7 weeks (Haas, 2018; Haas et al., 2012). These regimens aim to increase the local control probability as compared to surgery alone. Classically, in STS, tumor size, location, depth, and the French Federation of Cancer Centers Sarcoma Group histologic grading system (based on tumor differentiation, tumor necrosis, and mitotic activity) are the most important prognostic factors (Sbaraglia and Dei Tos, 2019). As well as stage, surgery, and preoperative RT could be crucial in achieving personalized treatment (Soydemir et al., 2020; Gao et al., 2021).

Recently, Artificial Intelligence (AI)-based solutions paved the way for the development of automatic and weird classification solutions. Such progress has been implemented in the field of imaging with the possibility of characterizing human tumors through "radiomics" texture analyses, which are based on several image-derived, quantitative measurements, including intensity histogram, spatial distribution relationships, and textural heterogeneity. This approach can be used to understand

the relationships between histological and imaging characteristics of STS, such as heterogeneity and their biological characteristics or expected prognosis and treatment outcomes. Radiomic features extracted from Magnetic Resonance (MR) images helped to distinguish low-grade from high-grade sarcomas and showed promise as biomarkers for predicting overall survival in patients with STS (Vallières et al., 2015; Crombé et al., 2019; Malinauskaite et al., 2020). The value of radiomics have also been assessed for differentiating STS of different histopathologic grades to enhance the precision of preoperative diagnosis (Xu et al., 2020). Furthermore, machine-learning model based on radiomics turned out favorable for preoperative prediction of distant metastasis from soft-tissue sarcoma to guide treatment strategies (Tian et al., 2021). More recently, Deep Learning (DL) techniques have also been employed in the field of STS. The performance of a DL radiomic nomogram has been evaluated as features extractor to predicting tumor relapse in patients with STS (Liu et al., 2021). Otherwise, DL has been used to accurately diagnose frequent subtypes of STS from conventional histopathological slides for diagnosis and prognosis survival prediction (Foersch et al., 2021). DL-based imaging analysis has also been applied as an alternative way to characterize and classify STS and to predict tumor grading (Navarro et al., 2021).

The main power of DL lies in its deep architecture using a hierarchical learning approach (Szegedy et al., 2015; Zeiler and Fergus, 2014), which allows for extracting a set of discriminating features at multiple levels of abstraction. Feature maps in the earlier layers extract low-level features (i.e. edges, shape, and textures), while feature maps in higher layers extract high-level features (i.e. abstract domain representation). However, training a Deep Neural Network (DNN) from scratch is a challenging task. It can be laborious and time-consuming, demanding a great deal of expertise and costly. In addition to the fact that it requires a large amount of labeled training data; a requirement that may be difficult to meet in the medical domain (Swati et al., 2019). Effectively, datasets in radiation oncology and medical physics tend to be small in size. They are generally in the hundreds and rarely in the thousands or millions of images. Furthermore, the extremely large number of parameters and hyperparameters that need to be tuned to train Neural Network (NN) raises the question of whether DL algorithms are truly appropriate in the field of medical images. In fact, it is known that data-driven approaches, especially DL, have difficulty achieving high performance on limited data sets which consequently limits the stability and generalizability of the model.

For the small dataset scenario, the use of pre-trained networks based on transfer learning and fine-tuning strategy is ubiquitous (Shin et al., 2016, 2017; Tajbakhsh et al., 2016). Transfer learning is a machine learning paradigm that aims at improving

the prediction performance of a learning task by applying knowledge previously gained in a related learning task (Zhu et al., 2022). Evidently, the extent of training that can be performed in the target task depends on the size of the target dataset, since the algorithms can easily overfit in small datasets (Yosinski et al., 2014). Generally speaking, exploring the way to build effective models based on small data by transferring information from auxiliary data meaningfully is an urgent task.

In this work, we sought to develop a deep learning model for predicting patient outcome in response to RT, in terms of distant metastasis development. Estimating and predicting treatment effects, especially during the treatment, would be valuable in monitoring patients' response to treatment, and hence provide a window for personalized treatment adaptation which enables improved treatment efficacy or reduced normal tissue complications. Several distinctive characteristics of our technique include $T1$-weighted, $T2$-weighted, contrast-enhanced $T1$-weighted MRI, Computed Tomography (CT) images and dose distributions and transfer learning approach. The main novelty concerns the features extraction from different combination of the imaging modalities which can lead to improved prediction performance.

7.2 Materials and Methods

Patient cohort

All patients enrolled in this retrospective study received pre-operative radiation therapy for soft tissue sarcomas delivered with image-guided intensity-modulated radiation therapy technique. In this single-center analysis, data of 72 patients from January 2011 to August 2020 were initially analyzed for further inclusion. The inclusion criteria were (i) availability of post-RT MRIs with diagnostic-quality throughout the follow-up period, (ii) availability of multi-parametric MRI, including axial $T1$-weighted, $T2$-weighted, Contrast-Enhanced (CE) $T1$-weighted maps, (iii) availability of radiotherapy CT, structures set, plan and $3D$ dose volume. Follow-up data were acquired by medical records and allowed to identify the development of distant metastasis as the surrogate of patient's response to radiotherapy treatment. Patients with incomplete clinical data, poor tumor tissue quality, and incomplete or poor-quality multimodal images were excluded from the research. The study population included 61 patients with a complete data set.

Imaging preprocessing

The proposed patient classification is based on two-dimensional images ($2D$ slices), not three-dimensional ($3D$) volume, because in most clinical practice, the acquired and available images are $2D$ slices with a large slice gape. Therefore, our classification system based on $2D$ images for clinical application is practical. Moreover, obtaining transversal $2D$ slices from $3D$ volume for each patient allowed us to increase the number of training samples for the deep neural networks. This means that from every patient in the training set we can generate as many training samples as transversal slices are available from the patient tumor. When counting the overall number of training samples, we can then go from several tens in the original dataset to thousands after slicing the patient.

Two fundamental data pre-processing steps was carried out across all the patients: images co-registration and resolution adjustment. $3D$ tumor contours propagation was obtained free from radiotherapy process by co-registering MR images on the centering CTs and applying deformable registration. In this way it was possible to obtain tumor segmentation also on MR images.

Since the acquisition volume of clinical images is much larger than the tumor region, we opted to extract only the tumor volume with a margin of 1 cm both superiorly and inferiorly in the cranio-caudal direction (z-axis). Furthermore, for MRI it is usual to have a reduced Field Of View (FOV) compared to CT to reduce acquisition times, therefore the axial matrices (on the $x - y$ plane) have a smaller dimension than those of CT in terms of number of pixels. For this reason, we resized all the multimodal images in 512×512 matrices, where pixel value outside the perimeter of the smaller images was set to 0. Consequently, we dealt with resampling along the z-axis, specifically at 3 mm slices, to obtain the same number of slices for the volume of interest for each type of multimodal image (k slices in the CT, k slices in the dose distribution and k slices in MRIs). The pre-trained network we considered in our study, requires input images of 224×224 pixels with RGB 3-channel, so the images were further scaled on the $x - y$ plane to fit the required input size. This double step was done to maintain the aspect ratio within the images.

Let's shift attention to another fundamental aspect. Intensity values in multimodal imaging do not have a fixed meaning and they vary greatly across CT, dose and MR images. Data mining and especially deep neural network approaches need to normalize the inputs; otherwise, the network will be ill-conditioned. In principle, normalization is performed to obtain the same range of values for each input into the NN model, which can guarantee a stable convergence. The intensity normalization brings the intensity values within a coherent range across all the multimodal images

and facilitates learning in the training process. In this scenario, normalization was performed employing min-max normalization to scale the image intensity values between 0 and 1, which is computed as follows:

$$y_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{7}$$

where y_i is the normalized intensity value against position x_i and $min(x)$ and $max(x)$ is the minimum and maximum intensity values, respectively, across the entire image.

A further step was necessary before obtaining the ideal database for training the neural network. It pertains to the three-channel tensor input. Since we have 5 different types of images available (CT, dose distribution, $T1w$, $T2w$ and CE $T1w$) and only 3 channels in the input, it was possible to create 10 different combinations (Figure 7.1), specifically: 1) $CT - Dose - T1w$, 2) $CT - Dose-$ CE $T1w$, 3) $CT - Dose - T2w$, 4) $CT - T1w-$ CE $T1w$, 5) $CT - T1w - T2w$, 6) $CT - T2w-$ CE $T1w$, 7) $Dose - T1w-$ CE $T1w$, 8) $Dose - T1w - T2w$, 9) $Dose - T2w-$ CE $T1w$, 10) $T1w - T2w-$ CE $T1w$. These combinations can be interpreted as 10 different sets (hereinafter referred to as 'combination sets') on which it is possible to train the neural network and considering the transversal $2D$ tumor slicing mentioned at the beginning of this section, each set counts 2395 image samples, all resized to $224 \times 224 \times 3$ according to the requirements of the pre-trained architecture for the $2D$ model.

Transfer learning and fine-tuning

The treatment response prediction network was constructed based on the deep convolutional network $VGG-19$ (Simonyan and Zisserman, 2014). It consists of sixteen convolutional layers and three fully connected layers (Figure 7.2). Our strategy is to apply transfer learning and therefore, we initialized weights from the ImageNet pre-trained $VGG-19$ model (Russakovsky et al., 2015) and fine-tuned on the ten combination sets one by one. Prior to this, the last fully connected layer was adapted to our binary classification task, as illustrated in Figure 7.2. The DNN was trained with fine-tuning the final block, made up of the last three fully connected layers, and keeping the weights of all other layers frozen.

Training Deep Learning Model

For the $VGG-19$ prediction network, the $SGDM$ optimizer was used with momentum at 0.9, mini-batch size of 64, an initial learning rate of 0.01 for 50 epochs. Training process was validated after every epoch and early stopping of the training

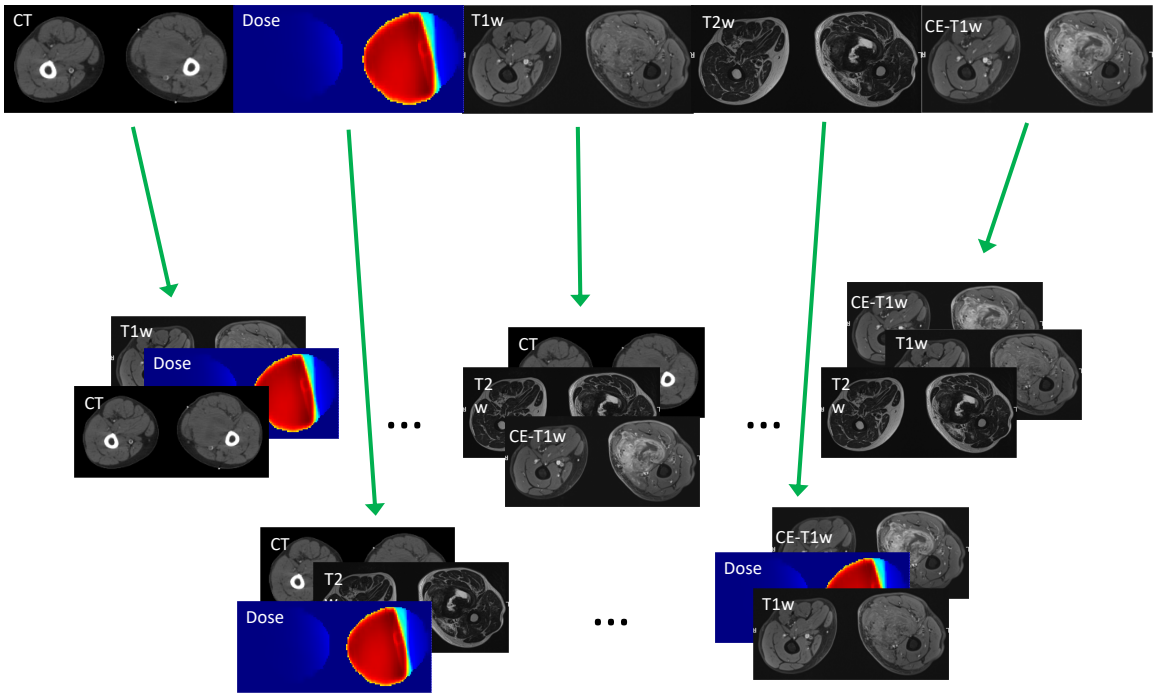


Figure 7.1: Different combinations of the multimodal images used as 3-channel input of the neural network. The figure shows only five of the ten possible combination sets.

was considered if the validation loss did not improve for 8 sequential epochs. No data augmentation was applied. The training was implemented with MATLAB (MATLAB 2020*b*, The MathWorks, Inc., Natick, Massachusetts, United States) and Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.6 GHz (28 CPUs) with 128GB RAM.

To test the performance of the proposed approach, for each combination set, we randomly divided the 2395 image samples into ten subsets of equal size such that each patient had been tested at least once during the independent testing stage. We ensured the no overlap and balanced classes across the ten subgroups. We exploited ten-fold cross-validation to evaluate the response prediction process. The final result is the average classification performance of the ten-fold test dataset.

Performance metrics

The prediction model was slice-based, where the three single-slice assembled by combining multimodal images in different ways were fed into the VGG prediction network as three channels. Accuracy, sensitivity (or recall), specificity, precision and $F1$-score metrics were reported (Gao et al., 2020; Zhu et al., 2022).

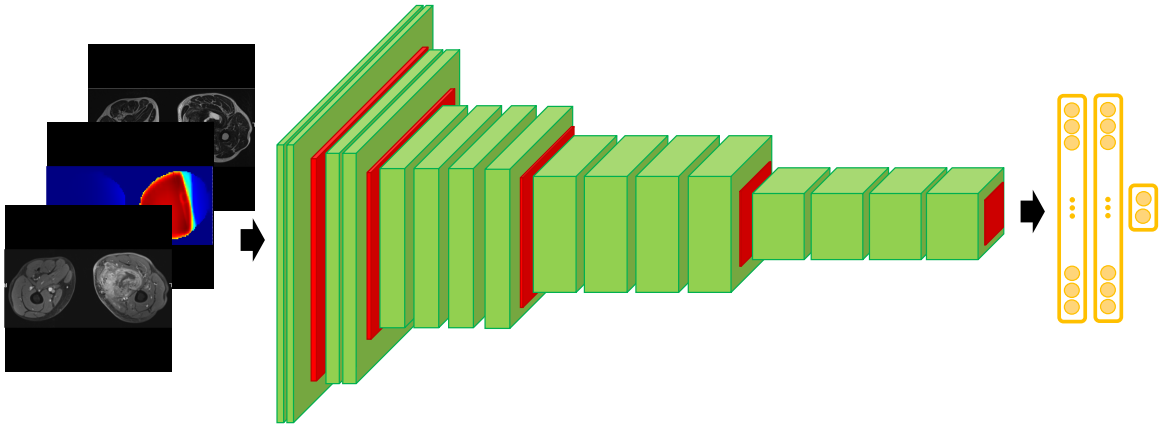


Figure 7.2: Architecture of the modified $VGG-19$ network for treatment response prediction; final layer adapted to our binary classification task. In green convolutional layers, in red pooling layers, in orange fully connected layers. Deep learning strategy: the network receives the $2D$ transversal slices and outputs the probability of the image for the two classes.

7.3 Results

The slice-based average prediction accuracy on training, validation, and test sets over the ten combination sets are shown in Figure 7.3. The benefit of transfer learning and fine-tuning is to reduce overfitting and speed the convergence. This benefit is very clear from the accuracy and the loss history of our proposed model (Figure 7.4). Validation and training losses were reduced over the epochs and converged with small gap between them in the plateau region, and at the same time training and validation accuracies reached consistently their maximum performance very fast.

Regarding the best performance among ten-fold cross-validation, the averaged accuracy was 0.93 ± 0.02, 0.92 ± 0.02, and 0.90 ± 0.03 for training, validation, and test respectively. These results confirmed the validity of the transfer learning with fine-tuning strategy to classify medical images. The prediction accuracy on the unseen test set was close to that on the validation set. A mild accuracy drop of 1% was observed. Overall, the first combination set had the worst performance on the test set while combination set 9 had the best prediction.

Table 7.1 depicts the classification metrics on the test sets over the ten-fold cross-validation for the ten combination sets. Overall, there is no clear difference between sensitivity and specificity, with an average value over the ten final models of 0.75 ± 0.06 and 0.78 ± 0.05, respectively. All models showed good precision of at least 0.68. Combination set 9 classified with the best accuracy of 0.90 ± 0.03. This is also reflected by the best sensitivity value of 0.85 ± 0.06 and with a good specificity of 0.84 ± 0.04. Combination set 10 has the best specificity and precision of 0.86 ± 0.02

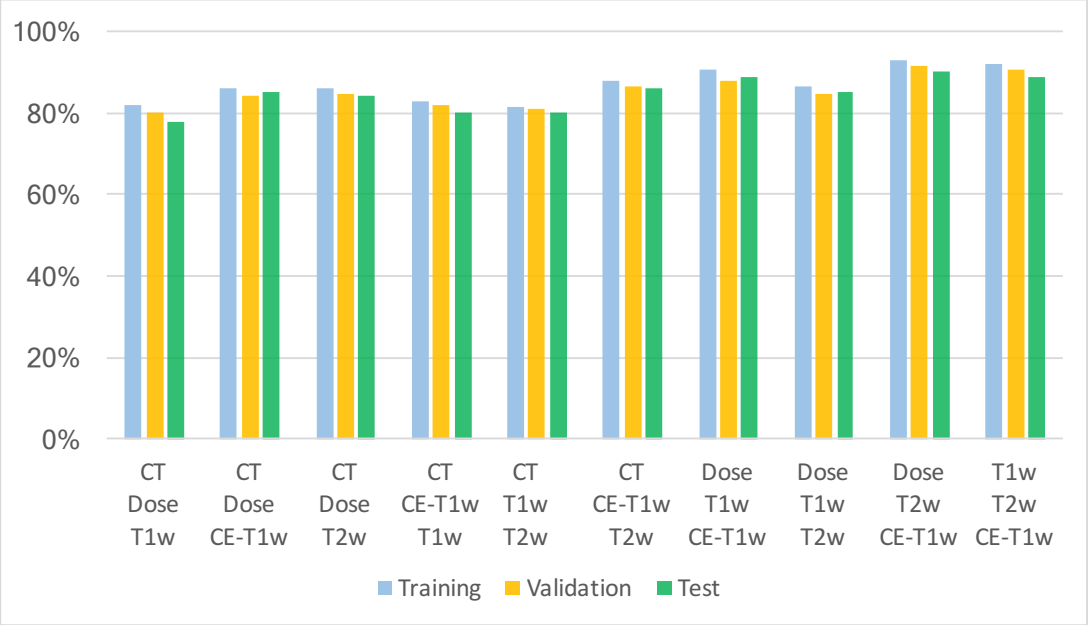


Figure 7.3: Slice-based ten-fold average training, validation and test accuracy results over the ten combination sets.

and 0.83 ± 0.03 but with the cost of a worse sensitivity value of 0.77 ± 0.02, leading to a total accuracy of 0.89 ± 0.02. In terms of the less imbalance-biased metric, $F1$-score, $Dose - T2w -$ CE $T1w$ set achieved the best result 0.81 ± 0.04.

7.4 Discussion

In this work, we sought to predict post-RT treatment effect, identified in the development of distant metastasis, of patients affected by STS using different combination of multimodal images. Empirical results have been shown that the generalization power of the deep networks is more than the shallow networks (Zhang et al., 2021a). To achieve effective training and consequently improved test performance of the deep neural networks, large datasets is often required. However, it is challenging to access and acquire a large size of data in the medical field. To resolve the small data size issue, the learning procedure was carried out in a transfer learning fashion with fine-tuning the last fully connected layers, adapted to our binary classification task, of a pre-trained $VGG - 19$ network. Transversal $2D$ slices from $3D$ tumor volume for each patient were employed to train and validate the prediction network, which was then tested on unseen patient data. The whole process was repeated ten times to evaluate the stability of the proposed workflow for each of the ten combination sets.

High accuracies were achieved on the training, validation, and test datasets. The best average accuracy was 0.90 ± 0.03 on the independent test sets for slice-based

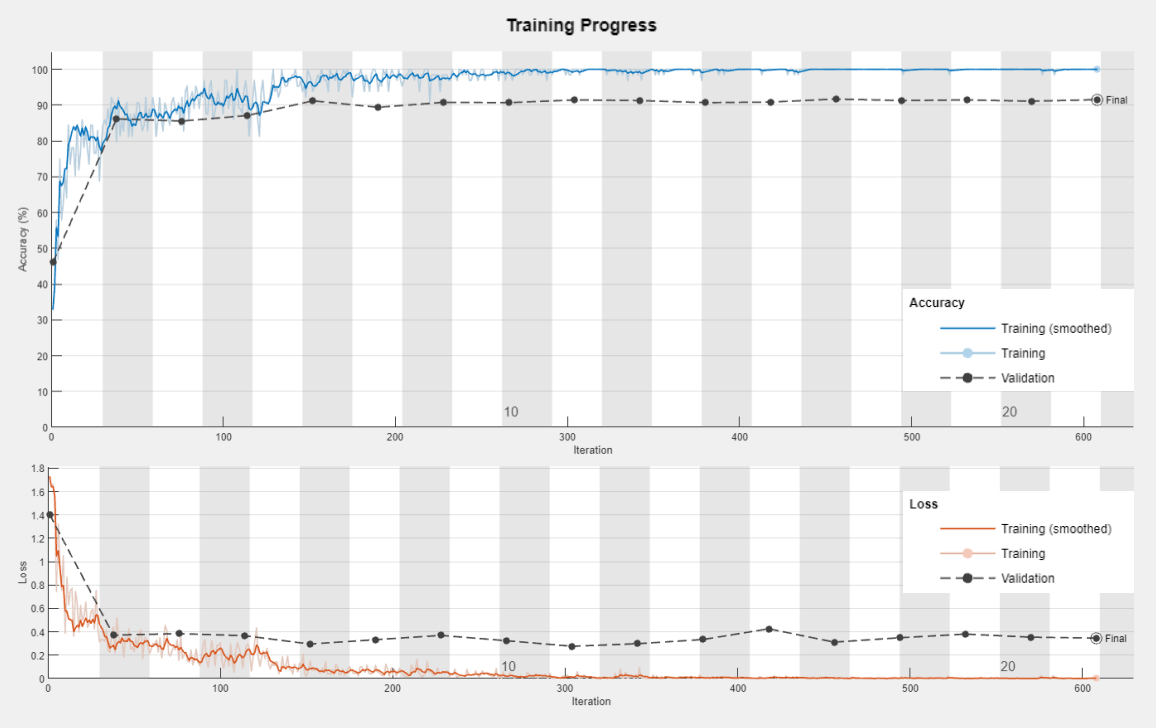


Figure 7.4: Exemplary learning curves for the $VGG-19$ classification network in an arbitrarily selected fold.

prediction considering $Dose-T2w-$ CE $T1w$ multimodal images combination. To the best of our knowledge, this is the first work that combines different MRI maps, CT and dose distribution with deep learning to predict treatment effect for sarcoma patients. Despite the result is still being preliminary, it demonstrated the feasibility and the efficacy of the proposed workflow on predicting post-RT patient outcome.

It must be emphasized that the dose distributions may provide additional information for the correct stratification of soft tissue sarcoma patients. In this perspective, we believe that it would be appropriate to develop more studies to better understand the role of this imaging biomarker in the assessment of patient outcome in STS. Analyzing the results shown in Table 7.1, it can be seen that the combinations in which the CT and $T1w$ images are present give the least performing results. While for the combinations in which the other three types of multimodal images are present ($Dose-T2w-$ CE $T1w$) the results are better. A possible explanation could be based on the greater information content present in these last images; it is no coincidence that the use of CE $T1w$ sequences serves precisely to make the heterogeneity of the lesion more evident. In addition, the great power of the DNN to capture the different heterogeneities and microstructures present within the lesson, as well as the ability to obtain a high level of abstraction thanks to the depth of the network itself, have allowed to obtain satisfactory results.

Table 7.1: Average classification performances of the proposed method over ten independent test sets for the ten combination sets. In bold, the best result for each metric is marked.

Combination set	Accuracy	Sensitivity	Specificity	Precision	F1-score
1. CT – Dose – T1w	0.78 ± 0.02	0.67 ± 0.02	0.75 ± 0.02	0.73 ± 0.05	0.70 ± 0.03
2. CT – Dose – CE-T1w	0.85 ± 0.03	0.81 ± 0.03	0.82 ± 0.03	0.77 ± 0.06	0.79 ± 0.04
3. CT – Dose – T2w	0.84 ± 0.03	0.75 ± 0.04	0.83 ± 0.03	0.79 ± 0.06	0.77 ± 0.05
4. CT – T1w – CE-T1w	0.80 ± 0.02	0.68 ± 0.02	0.72 ± 0.02	0.69 ± 0.05	0.68 ± 0.04
5. CT – T1w – T2w	0.80 ± 0.02	0.71 ± 0.03	0.70 ± 0.02	0.68 ± 0.04	0.69 ± 0.03
6. CT – T2w – CE-T1w	0.86 ± 0.04	0.81 ± 0.06	0.73 ± 0.04	0.71 ± 0.06	0.76 ± 0.06
7. Dose – T1w – CE-T1w	0.89 ± 0.04	0.75 ± 0.05	0.84 ± 0.04	0.81 ± 0.07	0.78 ± 0.06
8. Dose – T1w – T2w	0.85 ± 0.02	0.76 ± 0.03	0.80 ± 0.02	0.77 ± 0.05	0.76 ± 0.04
9. Dose – T2w – CE-T1w	**0.90 ± 0.03**	**0.85 ± 0.06**	0.84 ± 0.04	0.78 ± 0.05	**0.81 ± 0.04**
10. T1w – T2w – CE-T1w	0.89 ± 0.02	0.77 ± 0.02	**0.86 ± 0.02**	**0.83 ± 0.03**	0.80 ± 0.03

Early patient outcome prediction holds the promise of achieving personalized patient management. The capability of assessing and predicting their response to RT provides the opportunity for personalized treatment adaptation: conservative surgery or even avoidance of surgery may be adopted in patients with complete response to RT, whereas radiation boosting to the non-responding region might be beneficial for improving overall treatment efficacy in patients with insufficient response to RT.

There are a few limitations of this work. The main limitation of this work is the small patient cohort, and it has led to two consequences. First, the prediction model is built to predict treatment response for each slice whereas only one single score is available for each patient. We have implicitly assumed that all tumor imaging slices had the same score, so the data size is sufficient to allow the use of deep learning-based prediction. Second, repetitive cross-validation was applied to estimate the robustness and the results are potentially biased because we lack an external independent test set. We are enrolling more patients on an expansion cohort to improve the model robustness and, hopefully, to have enough data to formulate patient-based prediction. The second limitation relates to tumor heterogeneity. Although patients enrolled in this study each received the same treatment scheme, their histology subtypes are different. Sarcoma is known for its histologic and biologic diversity, and these differences are often reflected in divergent imaging characteristics. Lastly, the hardware at our disposal allowed us to perform a shallow fine-tuning in fact

only the final $VGG-19$ block was fine-tuned. We would also like to engage a deep fine-tuning, as it has been demonstrated that deep fine-tuning CNNs (preferably in a layer-wise manner) is useful for medical image analysis, performing as well as fully trained CNNs and even outperforming the latter when limited training data are available (Tajbakhsh et al., 2016).

In future studies, because in DL the trained model is hard to interpret, we plan to exploit Gradient Weighted Class Activation Map (Grad-CAM) as visual interpretation, which can shed light on the region of interest inside the original images that have a significant contribution to the final classification score. Moreover, we want to highlight how the combination of multimodal images is more statistically informative than examining only one type of image at a time. We would therefore like to repeat the experiment taking into account each type of images, one at a time, and verify that the performances obtained in this way are lower than the results presented in this study.

7.5 Conclusion

Pre-trained networks based on transfer learning and fine-tuning strategy possess important characteristics that make them natural candidates when applying deep learning to medical image tasks. Based on the results obtained in the present work, we believe that pretrained networks can be a very useful and powerful tool.

Because distant metastasis carries a poor prognosis, preoperative understanding of their likelihood in soft tissue sarcoma is clinically important. This preliminary study showed the feasibility of combining transfer learning and multimodal images to predict distant metastasis development after radiotherapy for soft tissue sarcoma patients. MR and dose maps turned out to be the best for our purpose. The model demonstrates good prognostic accuracy and provides a non-invasive opportunity for personalized treatment adaptation for improved treatment efficacy.

Chapter 8

Conclusion

This chapter summarizes the contribution of this thesis and discusses directions for future research. The thesis deals with the application of Artificial Intelligence models to support the clinical decision-making process with special focus on small medical imaging database. Though medical imaging has seen a growing interest in AI research, training models require a large amount of quality 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 ploys to fight overfitting otherwise they risk getting overconfident estimates, undermining the reliability of AI models. In addition, there are other challenges and pitfalls that undermine the success of a training and the achievement of a reproducible and generalizable AI model. Adequate curation, analysis, labeling, class imbalance, data leakage, external independent test, ethical issues and costs are detrimental for AI performance and critical to achieving high-impact clinically meaningful AI algorithms, if not properly accounted for. Executing AI pipelines on medical imaging data requires considerable effort in data curation and preprocessing. In effect, I was able to experience this difficulty at the forefront given that in the research projects presented in this thesis I followed and took care the creation of three medical images datasets for the AI algorithms development. In carrying out this task I have adhered as much as possible to the eight concepts introduced in chapter 2: ethical approval, data access, querying data, data de-identification, data transfer, quality control, structured data and label data.

The literature review was performed to identify the challenges and pitfalls of employing machine learning systems on small datasets of medical images. Transfer learning and data augmentation, especially GAN, could represent the most reasonable choices to fight overfitting. The results are promising and, in some studies, even too much (probably due to some kind of methodological bias). They provided a

proof-of-concept of the developed AI models, but they are not yet mature enough for large-scale implemented in the clinical setting and widely used. For studies on pediatric medulloblastoma and Soft-Tissue-Sarcoma, machine learning techniques return reasonable results and may prove a valuable and cost-effective aid by providing non-invasive quantitative data that integrate qualitative image information already available. Furthermore, it has been seen how the use of information extracted from dose distributions improves the performance of the algorithms. AI techniques offers efficient computational tools to disentangle the complexity of the information extracted from medical images and to discover sub-visual features that cannot be detected with traditional methods. Moreover, AI can provide a fully data-driven approach that learns to extract high-level features, avoiding the need for predefined instructions that might introduce cognitive biases. In particular, deep learning improves the efficiency and accuracy of statistical methods to analyze high-throughput imaging data. The results showed that a higher level of abstraction possibly can reveal unknown relationships within the data by describing, understanding and recognizing disease patterns. In this regard, pre-trained networks based on transfer learning strategy with fine-tuning suggests an acceptable and plausible approach to develop classification system. For the $COVID-19$ segmentation challenge, the lack of standardization is a major issue. Combining different public databases, each of them not largely populated, various lesion labeling and data selection criteria impacted in a relevant way on the performances. Processing aggregated data from different sources can be difficult if labeling has been performed using different guidelines. Further, the metadata contained in the DICOM header of medical images are very important in order to better manage the images acquired from different scanners, especially for data homogenization. Their full or partial absence due to the conversion to another format to allow publication of the database was an issue.

Despite the good performances obtained so far, there is still a lot of work to be done. Some strategies can be useful to implement AI in small medical imaging database such as transfer learning and data augmentation, but still remain the issue of the external validation of the models, using data that are independent from those of the training. It is an essential step to guarantee the reproducibility, the generalizability and the reliability of the developed AI algorithms. In this regard, due to the difficulties encountered in the data sharing, it is better to change point of view and taking the opposite path, that is to share the code or develop the distributed learning. In addition, it has been seen that greater levels of abstraction by incorporating greater information content lead to better performance, but it remains necessary to be able to interpret the developed models. In this direction, systems to verify transparency, interpretability and explainability must be designed and necessarily integrated into

the AI workflow. This is particularly relevant in medical context where medical professionals should be able to understand how and why a machine-based decision has been made in order to trust the decision and augment their decision-making process. For this reason, Gradient Weighted Class Activation Map (Grad-CAM) will be exploited in the Soft-Tissue-Sarcoma project since it is still ongoing. They can shed light on the region of interest inside the original images that have a significant contribution to the final classification score.

Although there are limitations and drawbacks hidden around the corner, Artificial Intelligence algorithms can be developed employing small and real hospital database to fulfill specific clinical needs and to support clinical decision-making process, provided that proper precautions and cares are taken into account.

Appendix A

Publications

Journal papers

1. F. Lizzi, F. Brero, R. F. Cabini, M. E. Fantacci, **S. Piffer**, I. Postuma, L. Rinaldi, A. Retico. "Making Data Big for a Deep-learning Analysis: Aggregation of Public COVID-19 Datasets of Lung Computed Tomography Scans", *Proceedings of the 10th International Conference on Data Science, Technology and Applications* (*DATA 2021*), pages: 316–321, 2021.

2. F. Lizzi, A. Agosti, F. Brero, R. F. Cabini, M. E. Fantacci, S. Figini, A. Lascialfari, F. Laruina, P. Oliva, **S. Piffer**, I. Postuma, L. Rinaldi, C. Talamonti, A. Retico. "Quantification of pulmonary involvement in COVID-19 pneumonia by means of a cascade of two U-nets: training and assessment on multiple datasets using different annotation criteria", *International Journal of Computer Assisted Radiology and Surgery*, Volume 17, pages: 229–237, 2022.

Papers under review

1. **S. Piffer**, L. Ubaldi, S. Tangaro, A. Retico, C. Talamonti. "How to deal the small data issue in artificial intelligence within medical images classification: a systematic review", *Computer Methods and Programs in Biomedicine*

2. **S. Piffer**, D. Greto, M. Mortilla, A. Ciccarone, C. Talamonti. "Radiomic and dosiomic-based unsupervised clustering development for radio-induced neurotoxicity in pediatric medulloblastoma", *Neuro-Oncology*

Papers not yet submitted

1. **S. Piffer**, D. Greto, M. Loi, C. Talamonti. "Development and validation of deep learning soft-tissue-sarcoma distant metastasis prediction based on transfer learning and fine-tuning with MR, CT and dose images"

Conference

1. C. Talamonti, **S. Piffer**, D. Greto, M. Mangoni, A. Ciccarone, M. E. Fantacci, F. Fusi, P. Oliva, M. Mortilla, L. Livi, S. Pallotta, A. Retico. "Radiomic and dosiomic profiling of paediatric Medulloblastoma tumours treated with Intensity Modulated Radiation Therapy", 3^{rd} *European Congress of Medical Physics, 2020*

2. **S. Piffer**, L. Ubaldi, D. Greto, M. Mortilla, A. Ciccarone, A. Retico, C. Talamonti. "Unsupervised classification of pediatric medulloblastoma tumors based on features from magnetic resonance images and dose distributions", *XII CONGRESSO NAZIONALE AIRMM - Associazione Italiana Risonanza Magnetica in Medicina*, **Best poster presentation 2021**

Bibliography

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. *Artificial Intelligence in Medicine*, 121:102176.

Adams, H. J. A., Kwee, T. C., Yakar, D., Hope, M. D., and Kwee, R. M. (2020a). Chest ct imaging signature of coronavirus disease 2019 infection: In pursuit of the scientific evidence. *Chest*, 158(5):1885–1895.

Adams, H. J. A., Kwee, T. C., Yakar, D., Hope, M. D., and Kwee, R. M. (2020b). Systematic review and meta-analysis on the value of chest ct in the diagnosis of coronavirus disease (covid-19): Sol scientiae, illustra nos. *AJR Am J Roentgenol*, 215(6):1342–1350.

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(4):161.

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(12):e05652.

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(2):223.

Ahmed, K. B., 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(2):345.

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 Transactions on Medical Imaging*, 39(7):2385–2394.

Alruwaili, M. and Gouda, W. (2022). Automated breast cancer detection models based on transfer learning. *Sensors*, 22(3):876.

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(7):1590.

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. *Scientific Reports*, 11(1).

An, P., Xu, S., Harmon, S., Turkbey, E., Sanford, T., Amalou, A., Kassin, M., Varble, N., Blain, M., Anderson, V., et al. (2020). Ct images in covid-19 [data set]. *The Cancer Imaging Archive*, 10.

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. *Journal of Medical and Biological Engineering*, 40(3):462–469.

Archer, T. C., Mahoney, E. L., and Pomeroy, S. L. (2017). Medulloblastoma: molecular classification-based personal therapeutics. *Neurotherapeutics*, 14(2):265–273.

Aryanto, K., Oudkerk, M., and van Ooijen, P. (2015). Free dicom de-identification tools in clinical research: functioning and safety of patient privacy. *European radiology*, 25(12):3685–3695.

Avanzo, M., Trianni, A., Botta, F., Talamonti, C., Stasi, M., and Iori, M. (2021). Artificial intelligence and the medical physicist: welcome to the machine. *Applied Sciences*, 11(4):1691.

Avanzo, M., Wei, L., Stancanello, J., Vallieres, M., Rao, A., Morin, O., Mattonen, S. A., and El Naqa, I. (2020). Machine and deep learning methods for radiomics. *Medical physics*, 47(5):e185–e202.

Ayana, G., Park, J., Jeong, J.-W., and woon Choe, S. (2022). A novel multistage transfer learning for ultrasound breast cancer image classification. *Diagnostics*, 12(1):135.

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 Computer Science*, 7:e555.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., and Davatzikos, C. (2017). Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. the cancer imaging archive. *Nat Sci Data*, 4:170117.

Baydilli, Y. Y. and Atila, Ü. (2020). Classification of white blood cells using capsule networks. *Computerized Medical Imaging and Graphics*, 80:101699.

Bien, N., Rajpurkar, P., Ball, R. L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B. N., Yeom, K. W., Shpanskaya, K., Halabi, S., Zucker, E., Fanton, G., Amanatullah, D. F., Beaulieu, C. F., Riley, G. M., Stewart, R. J., Blankenberg, F. G., Larson, D. B., Jones, R. H., Langlotz, C. P., Ng, A. Y., and Lungren, M. P. (2018). Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Medicine*, 15(11):e1002699.

Bouffet, E. (2021). Management of high-risk medulloblastoma. *Neurochirurgie*, 67(1):61–68.

Bradski, G. (2000). The opencv library. *Dr. Dobb's Journal of Software Tools*.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424.

Brehaut, J. C., Stiell, I. G., and Graham, I. D. (2006). Will a new clinical decision rule be widely used? the case of the canadian c-spine rule. *Academic emergency medicine*, 13(4):413–420.

Buda, M., Maki, A., and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259.

Buizza, G., Paganelli, C., D'ippolito, E., Fontana, G., Molinelli, S., Preda, L., Riva, G., Iannalfi, A., Valvo, F., Orlandi, E., et al. (2021). Radiomics and dosiomics for predicting local control after carbon-ion radiotherapy in skull-base chordoma. *Cancers*, 13(2):339.

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. *Computer Methods and Programs in Biomedicine*, 220:106815.

Cai, L., Gao, J., and Zhao, D. (2020). A review of the application of deep learning in medical image classification and segmentation. *Annals of translational medicine*, 8(11).

Carotti, M., Salaffi, F., Sarzi-Puttini, P., Agostini, A., Borgheresi, A., Minorati, D., Galli, M., Marotto, D., and Giovagnoni, A. (2020). Chest ct features of coronavirus disease 2019 (covid-19) pneumonia: key points for radiologists. *Radiol Med*, 125(7):636–646.

Chang, F.-C., Wong, T.-T., Wu, K.-S., Lu, C.-F., Weng, T.-W., Liang, M.-L., Wu, C.-C., Guo, W. Y., Chen, C.-Y., and Hsieh, K. L.-C. (2021). Magnetic resonance radiomics features and prognosticators in different molecular subtypes of pediatric medulloblastoma. *Plos one*, 16(7):e0255500.

Chatterjee, A., Vallieres, M., et al. (2018). An empirical approach for avoiding false-positives when applying high-dimensional radiomics to small datasets. *IEEE Trans. Radiat. Plasma Med. Sci*.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., and Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563.

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(11):e0187336.

Chougrad, H., Zouaki, H., and Alheyane, O. (2018). Deep convolutional neural networks for breast cancer screening. *Computer Methods and Programs in Biomedicine*, 157:19–30.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., and Prior, F. (2013). The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057.

Cogan, T., Cogan, M., and Tamil, L. (2019). MAPGI: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning. *Computers in Biology and Medicine*, 111:103351.

Coindre, J. M., Terrier, P., Guillou, L., Le Doussal, V., Collin, F., Ranchère, D., Sastre, X., Vilain, M. O., Bonichon, F., and N'Guyen Bui, B. (2001). Predictive value of grade for metastasis development in the main histologic types of adult soft tissue sarcomas: a study of 1240 patients from the french federation of cancer centers sarcoma group. *Cancer*, 91(10):1914–1926.

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(jan07 4):g7594–g7594.

Corino, V. D., Montin, E., Messina, A., Casali, P. G., Gronchi, A., Marchianò, A., and Mainardi, L. T. (2018). Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions. *Journal of Magnetic Resonance Imaging*, 47(3):829–840.

Crombé, A., Périer, C., Kind, M., De Senneville, B. D., Le Loarer, F., Italiano, A., Buy, X., and Saut, O. (2019). T2-based mri delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy. *Journal of Magnetic Resonance Imaging*, 50(2):497–510.

Crombé, A., Roulleau-Dugage, M., and Italiano, A. (2022). The diagnosis, classification, and treatment of sarcoma in this era of artificial intelligence and immunotherapy. *Cancer Communications*.

Cui, S., Tseng, H.-H., Pakela, J., Ten Haken, R. K., and El Naqa, I. (2020). Introduction to machine and deep learning for medical physicists. *Medical physics*, 47(5):e127–e147.

Dai, Y., Gao, Y., and Liu, F. (2021). TransMed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8):1384.

Dawud, A. M., Yurtkan, K., and Oztoprak, H. (2019). Application of deep learning in neuroradiology: brain haemorrhage classification using transfer learning. *Computational Intelligence and Neuroscience*, 2019.

de Vos, B. D., Wolterink, J. M., Leiner, T., de Jong, P. A., Lessmann, N., and Išgum, I. (2019). Direct automatic coronary calcium scoring in cardiac and chest ct. *IEEE transactions on medical imaging*, 38(9):2127–2138.

Desai, S., Baghal, A., Wongsurawat, T., Al-Shukri, S., Gates, K., Farmer, P., Rutherford, M., Blake, G., Nolan, T., et al. (2020). Data from chest imaging with clinical and genomic correlates representing a rural covid-19 positive population [data set]. the cancer imaging archive (2020).

DeVries, T. and Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., et al. (2014). The autism brain

imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667.

Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., Lituiev, D., Copeland, T. P., Aboian, M. S., Mari Aparici, C., et al. (2019). A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain. *Radiology*, 290(2):456–464.

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. *Scientific Reports*, 10(1).

Durgut, R., Aydin, M. E., and Rakib, A. (2022). Transfer learning for operator selection: A reinforcement learning approach. *Algorithms*, 15(1):24.

El Emam, K. and Dankar, F. K. (2008). Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15(5):627–637.

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. *Frontiers in Human Neuroscience*, 16.

Fang, X., Kruger, U., Homayounieh, F., Chao, H., Zhang, J., Digumarthy, S. R., Arru, C. D., Kalra, M. K., and Yan, P. (2021). Association of ai quantified covid-19 chest ct and patient outcome. *Int J Comput Assist Radiol Surg*, 16(3):435–445.

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. *Computerized Medical Imaging and Graphics*, 90:101897.

Feng, Y., Zhang, L., and Yi, Z. (2017). Breast cancer cell nuclei classification in histopathology images using deep neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 13(2):179–191.

Foersch, S., Eckstein, M., Wagner, D.-C., Gach, F., Woerl, A.-C., Geiger, J., Glasner, C., Schelbert, S., Schulz, S., Porubsky, S., Kreft, A., Hartmann, A., Agaimy, A., and Roth, W. (2021). Deep learning for diagnosis and survival prediction in soft tissue sarcoma. *Annals of Oncology*, 32(9):1178–1187.

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H.,

Weissman, M. M., and Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120.

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. *Medical Physics*, 47(10):4895–4905.

Gao, Y., Ghodrati, V., Kalbasi, A., Fu, J., Ruan, D., Cao, M., Wang, C., Eilber, F. C., Bernthal, N., Bukata, S., Dry, S. M., Nelson, S. D., Kamrava, M., Lewis, J., Low, D. A., Steinberg, M., Hu, P., and Yang, Y. (2021). Prediction of soft tissue sarcoma response to radiotherapy using longitudinal diffusion MRI and a deep neural network with generative adversarial network-based data augmentation. *Medical Physics*, 48(6):3262–3372.

Gao, Y., Kalbasi, A., Hsu, W., Ruan, D., Fu, J., Shao, J., Cao, M., Wang, C., Eilber, F. C., Bernthal, N., Bukata, S., Dry, S. M., Nelson, S. D., Kamrava, M., Lewis, J., Low, D. A., Steinberg, M., Hu, P., and Yang, Y. (2020). Treatment effect prediction for sarcoma patients treated with preoperative radiotherapy using radiomics features from longitudinal diffusion-weighted MRIs. *Physics in Medicine and Biology*, 65(17):175006.

Gatidis, S., Scharpf, M., Martirosian, P., Bezrukov, I., Küstner, T., Hennenlotter, J., Kruck, S., Kaufmann, S., Schraml, C., la Fougère, C., Schwenzer, N. F., and Schmidt, H. (2015). Combined unsupervised-supervised classification of multiparametric PET/MRI data: application to prostate cancer. *NMR in Biomedicine*, 28(7):914–922.

Gheshlaghi, S. H., Kan, C. N. E., and Ye, D. H. (2021). Breast cancer histopathological image classification with adversarial image synthesis. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*. IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014a). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014b). Generative adversarial networks.

GrandChallenge (GrandChallenge website). Grandchallenge 2020. https://covid-segmentation.grand-challenge.org/COVID-19-20/.

Gülbay, M., Özbay, B. O., Mendi, B. A. R., Baştuğ, A., and Bodur, H. (2021). A ct radiomics analysis of covid-19-related ground-glass opacities and consolidation: Is it valuable in a differential diagnosis with other atypical pneumonias? *Plos one*, 16(3):e0246582.

Haas, R. L. (2018). Preoperative radiotherapy in soft tissue sarcoma: from general guidelines to personalized medicine. *Chinese clinical oncology*, 7(4).

Haas, R. L., DeLaney, T. F., O'Sullivan, B., Keus, R. B., Le Pechoux, C., Olmi, P., Poulsen, J.-P., Seddon, B., and Wang, D. (2012). Radiotherapy for management of extremity soft tissue sarcomas: why, when, and where? *International Journal of Radiation Oncology* Biology* Physics*, 84(3):572–580.

Haga, A., Takahashi, W., Aoki, S., Nawa, K., Yamashita, H., Abe, O., and Nakagawa, K. (2017). Classification of early stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: interobserver delineation variability analysis. *Radiological Physics and Technology*, 11(1):27–35.

Han, C., Rundo, L., Araki, R., Nagano, Y., Furukawa, Y., Mauri, G., Nakayama, H., and Hayashi, H. (2019). Combining noise-to-image and image-to-image gans: Brain mr image augmentation for tumor detection. *Ieee Access*, 7:156966–156977.

Han, S., Williamson, B. D., and Fong, Y. (2021). Improving random forest predictions in small datasets from two-phase sampling designs. *BMC Medical Informatics and Decision Making*, 21(1).

Harvey, H. and Glocker, B. (2019). A standardised approach for preparing imaging data for machine learning tasks in radiology. In *Artificial intelligence in medical imaging*, pages 61–72. Springer.

Hashemzehi, R., Mahdavi, S. J. S., Kheirabadi, M., and Kamel, S. R. (2021). Y-net: a reducing gaussian noise convolutional neural network for MRI brain tumor classification with NADE concatenation. *Biomedical Physics and Engineering Express*, 7(5):055006.

Hashimoto, D., Rosman, G., Rus, D., and Meireles, O. (2018). Machine learning in surgery: Promises and perils. *Ann. Surg*, 268:70–76.

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. *Journal of Medical Imaging*, 5(03):1.

Hertel, R. and Benlamri, R. (2021). COV-SNET: A deep learning model for x-ray-based COVID-19 classification. *Informatics in Medicine Unlocked*, 24:100620.

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(4):e025611.

Ho, T. K. K. and Gwak, J. (2022). Feature-level ensemble approach for COVID-19 detection using chest x-ray images. *PLOS ONE*, 17(7):e0268430.

Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., and Langs, G. (2020). Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *Eur Radiol Exp*, 4(1):50.

Homeyer, A., Geißler, C., Schwen, L. O., Zakrzewski, F., Evans, T., Strohmenger, K., Westphal, M., Bülow, R. D., Kargl, M., Karjauv, A., Munné-Bertran, I., Retzlaff, C. O., Romero-López, A., Sołtysiński, T., Plass, M., Carvalho, R., Steinbach, P., Lan, Y.-C., Bouteldja, N., Haber, D., Rojas-Carulla, M., Sadr, A. V., Kraft, M., Krüger, D., Fick, R., Lang, T., Boor, P., Müller, H., Hufnagl, P., and Zerbe, N. (2022). Recommendations on compiling test datasets for evaluating artificial intelligence solutions in pathology. *Modern Pathology*, 35(12):1759–1769.

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*, 8:149808–149824.

Hu, B., Yan, L.-F., Yang, Y., Yu, Y., Sun, Q., Zhang, J., Nan, H.-Y., Han, Y., Hu, Y.-C., Sun, Y.-Z., Xiao, G., Tian, Q., Yue, C., Feng, J.-H., Zhai, L.-H., Zhao, D., Cui, G.-B., Welch, V. L., Cornett, E. M., Urits, I., Viswanath, O., Varrassi, G., Kaye, A. D., and Wang, W. (2021). Classification of prostate transitional zone cancer and hyperplasia using deep transfer learning from disease-related images. *Cureus*.

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. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J., and Cao,

B. (2020). Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *Lancet*, 395(10223):497–506.

Huynh, B. Q., Li, H., and Giger, M. L. (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501.

Hwang, E. J., Park, S., Jin, K.-N., Im Kim, J., Choi, S. Y., Lee, J. H., Goo, J. M., Aum, J., Yim, J.-J., Cohen, J. G., et al. (2019). Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA network open*, 2(3):e191095–e191095.

Igrec, J. and Fuchsjäger, M. H. (2021). Imaging of bone sarcomas and soft-tissue sarcomas. In *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, volume 193, pages 1171–1182. Georg Thieme Verlag KG.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Ishiguro, T., Kobayashi, Y., Uozumi, R., Takata, N., Takaku, Y., Kagiyama, N., Kanauchi, T., Shimizu, Y., and Takayanagi, N. (2019). Viral pneumonia requiring differentiation from acute and progressive diffuse interstitial lung diseases. *Intern Med*, 58(24):3509–3519.

Jack, C. R. J., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L Whitwell, J., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L. G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C. S., Krueger, G., Ward, H. A., Metzger, G. J., Scott, K. T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J. P., Fleisher, A. S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., and Weiner, M. W. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. *J Magn Reson Imaging*, 27(4):685–691.

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. *Artificial Intelligence*, 304:103652.

JL, P. L. E. B. R. (2017). Little kj demirer m qian s white rd automated critical test findings identification and online notification system using artificial intelligence in imaging. *Radiology*, 285:923–931.

Kaur, T. and Gandhi, T. K. (2022). Classifier fusion for detection of COVID-19 from CT scans. *Circuits, Systems, and Signal Processing*, 41(6):3397–3414.

Keras (2015). Keras. https://keras.io.

Khan, H. A., , Jue, W., Mushtaq, M., Mushtaq, M. U., , and and (2020). Brain tumor classification in MRI image using convolutional neural network. *Mathematical Biosciences and Engineering*, 17(5):6203–6216.

Kickingereder, P., Götz, M., Muschelli, J., Wick, A., Neuberger, U., Shinohara, R. T., Sill, M., Nowosielski, M., Schlemmer, H.-P., Radbruch, A., et al. (2016). Large-scale radiomic profiling of recurrent glioblastoma identifies an imaging predictor for stratifying anti-angiogenic treatment responseradiomic profiling of bev efficacy in glioblastoma. *Clinical Cancer Research*, 22(23):5765–5771.

Kiser, K. J., Ahmed, S., Stieb, S., Mohamed, A. S. R., Elhalawani, H., Park, P. Y. S., Doyle, N. S., Wang, B. J., Barman, A., Li, Z., Zheng, W. J., Fuller, C. D., and Giancardo, L. (2020). PleThora: Pleural effusion and thoracic cavity segmentations in diseased lungs for benchmarking chest CT processing pipelines. *Medical Physics*, 47(11):5941–5952.

Kiser, K. J., Barman, A., Stieb, S., Fuller, C. D., and Giancardo, L. (2021). Novel autosegmentation spatial similarity metrics capture the time required to correct segmentations better than traditional metrics in a thoracic cavity segmentation workflow. *Journal of Digital Imaging*, 34(3):541–553.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273–324. Relevance.

Lambin, P. (2017). Leijenaar rth deist tm peerlings j de jong eec van timmeren j sanduleanu s larue rthm even ajg jochems a et al. *Radiomics: The bridge between medical imaging and personalized medicine Nat Rev Clinic Oncol*, 14(749):10–1038.

Langlotz, C. P., Allen, B., Erickson, B. J., Kalpathy-Cramer, J., Bigelow, K., Cook, T. S., Flanders, A. E., Lungren, M. P., Mendelson, D. S., Rudie, J. D., Wang, G., and Kandarpa, K. (2019). A roadmap for foundational research on artificial intelligence in medical imaging: From the 2018 NIH/RSNA/ACR/the academy workshop. *Radiology*, 291(3):781–791.

Lavasa, E., Giannopoulos, G., Papaioannou, A., Anastasiadis, A., Daglis, I., Aran, A., Pacheco, D., and Sanahuja, B. (2021). Assessing the predictability of solar energetic particles with the use of machine learning techniques. *Solar Physics*, 296(7):1–47.

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. *Scientific Reports*, 12(1).

Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., and Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9.

Lehman, C. D., Yala, A., Schuster, T., Dontchos, B., Bahl, M., Swanson, K., and Barzilay, R. (2019). Mammographic breast density assessment using deep learning: clinical implementation. *Radiology*, 290(1):52–58.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563.

Lessmann, N., Sánchez, C. I., Beenen, L., Boulogne, L. H., Brink, M., Calli, E., Charbonnier, J.-P., Dofferhoff, T., van Everdingen, W. M., Gerke, P. K., et al. (2021). Automated assessment of covid-19 reporting and data system and chest ct severity scores in patients suspected of having covid-19 using artificial intelligence. *Radiology*, 298(1):E18–E28.

Levine, A. B., Peng, J., Farnell, D., Nursey, M., Wang, Y., Naso, J. R., Ren, H., Farahani, H., Chen, C., Chiu, D., Talhouk, A., Sheffield, B., Riazy, M., Ip, P. P., Parra-Herran, C., Mills, A., Singh, N., Tessier-Cloutier, B., Salisbury, T., Lee, J., Salcudean, T., Jones, S. J., Huntsman, D. G., Gilks, C. B., Yip, S., and Bashashati, A. (2020). Synthesis of diagnostic quality cancer pathology images.

Li, F., Chen, H., Liu, Z., dian Zhang, X., shan Jiang, M., zheng Wu, Z., and qian Zhou, K. (2019). Deep learning-based automated detection of retinal diseases using optical coherence tomography images. *Biomedical Optics Express*, 10(12):6204.

Lian, C., Ruan, S., Denœux, T., Jardin, F., and Vera, P. (2016). Selecting radiomic features from fdg-pet images for cancer treatment outcome prediction. *Medical image analysis*, 32:257–268.

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. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*. IEEE.

Liu, S., Shah, Z., Sav, A., Russo, C., Berkovsky, S., Qian, Y., Coiera, E., and Ieva, A. D. (2020). Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. *Scientific Reports*, 10(1).

Liu, S., Sun, W., Yang, S., Duan, L., Huang, C., Xu, J., Hou, F., Hao, D., Yu, T., and Wang, H. (2021). Deep learning radiomic nomogram to predict recurrence in soft tissue sarcoma: a multi-institutional study. *European Radiology*, 32(2):793–805.

Liu, Z.-m., Zhang, H., Ge, M., Hao, X.-L., An, X., and Tian, Y.-J. (2022). Radiomics signature for the prediction of progression-free survival and radiotherapeutic benefits in pediatric medulloblastoma. *Child's Nervous System*, 38(6):1085–1094.

Lorenzi, M., McMillan, A. J., Siegel, L. S., Zumbo, B. D., Glickman, V., Spinelli, J. J., Goddard, K. J., Pritchard, S. L., Rogers, P. C., and McBride, M. L. (2009). Educational outcomes among survivors of childhood cancer in british columbia, canada: report of the childhood/adolescent/young adult cancer survivors (cayacs) program. *Cancer*, 115(10):2234–2245.

Ma, J., Wang, Y., An, X., Ge, C., Yu, Z., Chen, J., Zhu, Q., Dong, G., He, J., He, Z., Cao, T., Zhu, Y., Nie, Z., and Yang, X. (2021a). Toward data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. *Med Phys*, 48(3):1197–1210.

Ma, J., Wang, Y., An, X., Ge, C., Yu, Z., Chen, J., Zhu, Q., Dong, G., He, J., He, Z., Nie, Z., and Yang, X. (2020). Towards efficient covid-19 ct annotation: A benchmark for lung and infection segmentation.

Ma, Y., Liu, J., Liu, Y., Fu, H., Hu, Y., Cheng, J., Qi, H., Wu, Y., Zhang, J., and Zhao, Y. (2021b). Structure and illumination constrained gan for medical image enhancement. *IEEE Transactions on Medical Imaging*, 40(12):3955–3967.

Mackie, T., Holmes, T., Swerdloffff, S., et al. (1993). A new concept for the delivery of conformal radiotherapy. *Med Phys*, 20(6):1709–1719.

Malinauskaite, I., Hofmeister, J., Burgermeister, S., Neroladaki, A., Hamard, M., Montet, X., and Boudabbous, S. (2020). Radiomics and machine learning differentiate soft-tissue lipoma and liposarcoma better than musculoskeletal radiologists. *Sarcoma*, 2020:1–9.

Mennes, M., Biswal, B. B., Castellanos, F. X., and Milham, M. P. (2013). Making data sharing work: the fcp/indi experience. *Neuroimage*, 82:683–691.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso, J. J., Criminisi, A., Das, T., Delingette, H., Demiralp, C., Durst, C. R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K. M., Jena, R., John, N. M., Konukoglu, E., Lashkari, D., Mariz, J. A., Meier, R., Pereira, S., Precup, D., Price, S. J., Raviv, T. R., Reza, S. M. S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.-C., Shotton, J., Silva, C. A., Sousa, N., Subbanna, N. K., Szekely, G., Taylor, T. J., Thomas, O. M., Tustison, N. J., Unal, G., Vasseur, F., Wintermark, M., Ye, D. H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., and Leemput, K. V. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024.

Mikołajczyk, A. and Grochowski, M. (2018). Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE.

Millard, N. E. and De Braganca, K. C. (2016). Medulloblastoma. *Journal of child neurology*, 31(12):1341–1353.

Moher, D., Liberati, A., Tetzlaff, J., and and, D. G. A. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*, 339(jul21 1):b2535–b2535.

Montoya, J., Li, Y., Strother, C., and Chen, G.-H. (2018). 3d deep learning angiography (3d-DLA) from c-arm conebeam CT. *American Journal of Neuroradiology*, 39(5):916–922.

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., 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. *Annals of Internal Medicine*, 162(1):W1–W73.

Moore, S. M., Maffitt, D. R., Smith, K. E., Kirby, J. S., Clark, K. W., Freymann, J. B., Vendt, B. A., Tarbox, L. R., and Prior, F. W. (2015). De-identification of medical images with retention of scientific research value. *Radiographics*, 35(3):727–735.

Morozov, S. P., Andreychenko, A., Pavlov, N., Vladzymyrskyy, A., Ledikhova, N., Gombolevskiy, V., Blokhin, I. A., Gelezhe, P., Gonchar, A., and Chernina, V. Y. (2020a). Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*.

Morozov, S. P., Andreychenko, A. E., Blokhin, I. A., Gelezhe, P. B., Gonchar, A. P., Nikolaev, A. E., Pavlov, N. A., Chernina, V. Y., and Gombolevskiy, V. A. (2020b). Mosmeddata: data set of 1110 chest ct scans performed during the covid-19 epidemic. *Digital Diagnostics*, 1(1):49–59.

MosMed (MosMed dataset website). Mosmed 2020. https://mosmed.ai/en/.

Moxon-Emre, I., Bouffet, E., Taylor, M. D., Laperriere, N., Scantlebury, N., Law, N., Spiegler, B. J., Malkin, D., Janzen, L., and Mabbott, D. (2014). Impact of craniospinal dose, boost volume, and neurologic complications on intellectual outcome in patients with medulloblastoma. *Journal of Clinical Oncology*, 32(17):1760–1768.

Muhammad, K., Ullah, H., Khan, Z. A., Saudagar, A. K. J., AlTameem, A., AlKhathami, M., Khan, M. B., Hasanat, M. H. A., Malik, K. M., Hijji, M., and Sajjad, M. (2022). WEENet: An intelligent system for diagnosing COVID-19 and lung cancer in IoMT environments. *Frontiers in Oncology*, 11.

Mutasa, S., Chang, P. D., Ruzal-Shapiro, C., and Ayyala, R. (2018). MABAL: a novel deep-learning architecture for machine-assisted bone age labeling. *Journal of Digital Imaging*, 31(4):513–519.

Mzoughi, H., Njeh, I., Wali, A., Slima, M. B., BenHamida, A., Mhiri, C., and Mahfoudhe, K. B. (2020). Deep multi-scale 3d convolutional neural network (CNN) for MRI gliomas brain tumor classification. *Journal of Digital Imaging*, 33(4):903–915.

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. *Biomedical Signal Processing and Control*, 68:102750.

Nalepa, J., Marcinkiewicz, M., and Kawulok, M. (2019). Data augmentation for brain-tumor segmentation: a review. *Frontiers in computational neuroscience*, 13:83.

Navarro, F., Dapper, H., Asadpour, R., Knebel, C., Spraker, M. B., Schwarze, V., Schaub, S. K., Mayr, N. A., Specht, K., Woodruff, H. C., Lambin, P., Gersing, A. S., Nyflot, M. J., Menze, B. H., Combs, S. E., and Peeken, J. C. (2021). Development and external validation of deep-learning-based tumor grading models in soft-tissue sarcoma patients using MR imaging. *Cancers*, 13(12):2866.

Nensa, F., Demircioglu, A., and Rischpler, C. (2019). Artificial intelligence in nuclear medicine. *Journal of Nuclear Medicine*, 60(Supplement 2):29S–37S.

Nichelli, L. and Casagranda, S. (2021). Current emerging mri tools for radionecrosis and pseudoprogression diagnosis. *Current Opinion in Oncology*, 33(6):597.

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. *Scientific Reports*, 7(1).

Organization, W. H. (2020). Global surveillance for covid-19 caused by human infection with covid-19 virus: interim guidance, 20 march 2020. Technical documents, World Health Organization.

Ovalle-Magallanes, E., Avina-Cervantes, J. G., Cruz-Aceves, I., and Ruiz-Pinales, J. (2020). Transfer learning for stenosis detection in x-ray coronary angiography. *Mathematics*, 8(9):1510.

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. *Journal of Medical Internet Research*, 22(11):e18563.

Packer, R. J., Zhou, T., Holmes, E., Vezina, G., and Gajjar, A. (2013). Survival and secondary tumors in children with medulloblastoma receiving radiotherapy and adjuvant chemotherapy: results of children's oncology group trial a9961. *Neuro-oncology*, 15(1):97–103.

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 Transactions*, 128:1–10.

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*, 23:125–132.

Parakh, A., Lee, H., Lee, J. H., Eisner, B. H., Sahani, D. V., and Do, S. (2019). Urinary stone detection on ct images using deep convolutional neural networks: evaluation of model performance and generalization. *Radiology. Artificial intelligence*, 1(4).

Park, S. H. and Han, K. (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, 286(3):800–809.

Parmar, C., Grossmann, P., Rietveld, D., Rietbergen, M., Lambin, P., and Aerts, H. (2015). Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. front oncol. 2015; 5: 272.

Parsons, D. W., Li, M., Zhang, X., Jones, S., Leary, R. J., Lin, J. C.-H., Boca, S. M., Carter, H., Samayoa, J., Bettegowda, C., Gallia, G. L., Jallo, G. I., Binder, Z. A., Nikolsky, Y., Hartigan, J., Smith, D. R., Gerhard, D. S., Fults, D. W., VandenBerg, S., Berger, M. S., Marie, S. K. N., Shinjo, S. M. O., Clara, C., Phillips, P. C., Minturn, J. E., Biegel, J. A., Judkins, A. R., Resnick, A. C., Storm, P. B., Curran, T., He, Y., Rasheed, B. A., Friedman, H. S., Keir, S. T., McLendon, R., Northcott, P. A., Taylor, M. D., Burger, P. C., Riggins, G. J., Karchin, R., Parmigiani, G., Bigner, D. D., Yan, H., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2011). The genetic landscape of the childhood cancer medulloblastoma. *Science*, 331(6016):435–439.

Perreault, S., Ramaswamy, V., Achrol, A., Chao, K., Liu, T., Shih, D., Remke, M., Schubert, S., Bouffet, E., Fisher, P., et al. (2014). Mri surrogates for molecular subgroups of medulloblastoma. *American Journal of Neuroradiology*, 35(7):1263–1269.

Phan, T. H. and Yamamoto, K. (2020). Resolving class imbalance in object detection with weighted cross entropy losses.

Ramaswamy, V. and Taylor, M. D. (2017). Medulloblastoma: from myth to molecular. *Journal of Clinical Oncology*, 35(21):2355–2363.

Rendón, E., Abundez, I., Arizmendi, A., and Quiroz, E. M. (2011). Internal versus external cluster validation indexes. *International Journal of Computers and Communications*.

Rivals, I., Personnaz, L., Taing, L., and Potier, M.-C. (2007). Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407.

Rocca, A., Brunese, M. C., Santone, A., Avella, P., Bianco, P., Scacchi, A., Scaglione, M., Bellifemine, F., Danzi, R., Varriano, G., Vallone, G., Calise, F., and Brunese, L. (2021). Early diagnosis of liver metastases from colorectal cancer through CT radiomics and formal methods: A pilot study. *Journal of Clinical Medicine*, 11(1):31.

Rodriguez, J. D., Perez, A., and Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):569–575.

Romero, M., Interian, Y., Solberg, T., and Valdes, G. (2020). Targeted transfer learning to improve performance in small medical physics datasets. *Medical Physics*, 47(12):6246–6256.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sahoo, P., Roy, I., Ahlawat, R., Irtiza, S., and Khan, L. (2021). Potential diagnosis of COVID-19 from chest x-ray and CT findings using semi-supervised learning. *Physical and Engineering Sciences in Medicine*, 45(1):31–42.

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. *Medical Physics*, 48(6):2827–2837.

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. *Physics in Medicine and Biology*, 65(10):105002.

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 Transactions on Medical Imaging*, 41(11):3278–3288.

Sbaraglia, M. and Dei Tos, A. P. (2019). The pathology of soft tissue sarcomas. *La radiologia medica*, 124(4):266–281.

Schelb, P., Kohl, S., Radtke, J. P., Wiesenfarth, M., Kickingereder, P., Bickelhaupt, S., Kuder, T. A., Stenzinger, A., Hohenfellner, M., Schlemmer, H.-P., et al. (2019). Classification of cancer at prostate mri: deep learning versus clinical pi-rads assessment. *Radiology*, 293(3):607–617.

Schneider, N., Strauss, D. C., Smith, M. J., Miah, A. B., Zaidi, S., Benson, C., van Houdt, W. J., Jones, R. L., Hayes, A. J., Fisher, C., et al. (2017). The adequacy of core biopsy in the assessment of smooth muscle neoplasms of soft tissues. *The American journal of surgical pathology*, 41(7):923–931.

Seidel, C., Heider, S., Hau, P., Glasow, A., Dietzsch, S., and Kortmann, R.-D. (2021). Radiotherapy in medulloblastoma—evolution of treatment, current concepts and future perspectives. *Cancers*, 13(23):5945.

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. *Translational Cancer Research*, 8(5):1741–1749.

Shen, C., Nguyen, D., Zhou, Z., Jiang, S. B., Dong, B., and Jia, X. (2020a). An introduction to deep learning in medical physics: advantages, potential, and challenges. *Physics in Medicine & Biology*, 65(5):05TR01.

Shen, Y., Li, X., Liang, X., Xu, H., Li, C., Yu, Y., and Qiu, B. (2020b). A deep-learning-based approach for adenoid hypertrophy diagnosis. *Medical Physics*, 47(5):2171–2181.

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. *Computer Methods and Programs in Biomedicine*, 196:105611.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2017). Three aspects on using convolutional neural networks for computer-aided detection in medical imaging. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing*, pages 113–136. Springer International Publishing.

Simard, P., Steinkraus, D., and Platt, J. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* IEEE Comput. Soc.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition.

Soffer, S., Ben-Cohen, A., Shimon, O., Amitai, M. M., Greenspan, H., and Klang, E. (2019). Convolutional neural networks for radiologic images: a radiologist's guide. *Radiology*, 290(3):590–606.

Sollini, M., Antunovic, L., Chiti, A., and Kirienko, M. (2019). Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. *European journal of nuclear medicine and molecular imaging*, 46(13):2656–2672.

Soydemir, G. P., Bahat, Z., Kandaz, M., Canyilmaz, E., Yöney, A., et al. (2020). Prognostic factors and clinical course of extremity soft-tissue sarcomas. *Journal of Cancer Research and Therapeutics*, 16(4):903.

Suganyadevi, S. and Seethalakshmi, V. (2022). CVD-HNet: Classifying pneumonia and COVID-19 in chest x-ray images using deep network. *Wireless Personal Communications*, 126(4):3279–3303.

Sun, R., Limkin, E. J., Vakalopoulou, M., Dercle, L., Champiat, S., Han, S. R., Verlingue, L., Brandao, D., Lancia, A., Ammari, S., Hollebecque, A., Scoazec, J.-Y., Marabelle, A., Massard, C., Soria, J.-C., Robert, C., Paragios, N., Deutsch, E., and Ferté, C. (2018). A radiomics approach to assess tumour-infiltrating cd8 cells and response to anti-pd-1 or anti-pd-l1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol*, 19(9):1180–1191.

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. *Computerized Medical Imaging and Graphics*, 75:34–46.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312.

Talha, I. and Hazrat, A. (2018). Generative adversarial network for medical images (mi-gan)[j]. *Journal of Medical Systems*, 42(11):231.

Tamayo, P., Cho, Y.-J., Tsherniak, A., Greulich, H., Ambrogio, L., Schouten-van Meeteren, N., Zhou, T., Buxton, A., Kool, M., Meyerson, M., et al. (2011). Predicting relapse in patients with medulloblastoma by integrating evidence from clinical and genomic features. *Journal of Clinical Oncology*, 29(11):1415.

Tian, L., Zhang, D., Bao, S., Nie, P., Hao, D., Liu, Y., Zhang, J., and Wang, H. (2021). Radiomics-based machine-learning method for prediction of distant metastasis from soft-tissue sarcomas. *Clinical Radiology*, 76(2):158.e19–158.e25.

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. *International Journal of Computer Assisted Radiology and Surgery*, 16(2):241–251.

Torgyn, S., Lowe, D., Daga, S., Briggs, D., Higgins, R., and Khovanova, N. (2015). Machine learning for predictive modelling based on small data. *Biomed. Eng*, 48:469–474.

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 Journal of Biomedical and Health Informatics*, 23(3):923–930.

Ubaldi, L., Valenti, V., Borgese, R., Collura, G., Fantacci, M., Ferrera, G., Iacoviello, G., Abbate, B., Laruina, F., Tripoli, A., Retico, A., and Marrale, M. (2021). Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples. *Physica Medica*, 90:13–22.

Uemura, T., Näppi, J. J., Ryu, Y., Watari, C., Kamiya, T., and Yoshida, H. (2020). A generative flow-based model for volumetric data augmentation in 3d deep learning for computed tomographic colonography. *International Journal of Computer Assisted Radiology and Surgery*, 16(1):81–89.

Usman, M., Zia, T., and Tariq, A. (2022). Analyzing transfer learning of vision transformers for interpreting chest radiography. *Journal of Digital Imaging*, 35(6):1445–1462.

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14(11):e0224365.

Vallières, M., Freeman, C. R., Skamene, S. R., and Naqa, I. E. (2015). A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Physics in Medicine and Biology*, 60(14):5471–5496.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K. (2013). The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79.

Van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G., Fillion-Robin, J.-C., Pieper, S., and Aerts, H. J. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107.

Von Bueren, A. O., Kortmann, R.-D., von Hoff, K., Friedrich, C., Mynarek, M., Müller, K., Goschzik, T., Zur Mühlen, A., Gerber, N., Warmuth-Metz, M., et al. (2016). Treatment of children and adolescents with metastatic medulloblastoma and prognostic relevance of clinical and biologic parameters. *Journal of clinical oncology*, 34(34):4151–4160.

Wang, C., Elazab, A., Wu, J., and Hu, Q. (2017a). Lung nodule classification using deep feature fusion in chest radiography. *Computerized Medical Imaging and Graphics*, 57:10–18.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017b). ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

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. *Quantitative Imaging in Medicine and Surgery*, 10(6):1249–1264.

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. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE.

Xi, P., Shu, C., and Goubran, R. (2018). Abnormality detection in mammography using deep convolutional neural networks. In *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE.

Xia, T., Kumar, A., Feng, D., and Kim, J. (2018). Patch-level tumor classification in digital histopathology images with domain adapted deep learning. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE.

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. *Frontiers in Oncology*, 10.

Xie, W., Jacobs, C., Charbonnier, J.-P., and Van Ginneken, B. (2020). Relational modeling for robust and efficient pulmonary lobe segmentation in ct scans. *IEEE transactions on medical imaging*, 39(8):2664–2675.

Xu, W., Hao, D., Hou, F., Zhang, D., and Wang, H. (2020). Soft tissue sarcoma: Preoperative MRI-based radiomics and machine learning may be accurate predictors of histopathologic grade. *American Journal of Roentgenology*, 215(4):963–969.

Xu, Y., Hu, M., Liu, H., Yang, H., Wang, H., Lu, S., Liang, T., Li, X., Xu, M., Li, L., Li, H., Ji, X., Wang, Z., Li, L., Weinreb, R. N., and Wang, N. (2021). A hierarchical deep learning approach with transparency and interpretability based on small samples for glaucoma diagnosis. *npj Digital Medicine*, 4(1).

Yala, A., Schuster, T., Miles, R., Barzilay, R., and Lehman, C. (2019). A deep learning model to triage screening mammograms: a simulation study. *Radiology*, 293(1):38–46.

Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629.

Yan, J., Zhang, S., Li, K. K.-W., Wang, W., Li, K., Duan, W., Yuan, B., Wang, L., Liu, L., Zhan, Y., et al. (2020). Incremental prognostic value and underlying biological pathways of radiomics patterns in medulloblastoma. *EBioMedicine*, 61:103093.

Yang, J., Sharp, G., Veeraraghavan, H., van Elmpt, W., Dekker, A., Lustberg, T., and Gooding, M. (2017). Data from lung ct segmentation challenge. *The cancer imaging archive*, 20.

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., and Shen, F. (2022). Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*.

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. *Computerized Medical Imaging and Graphics*, 80:101675.

Yi, P. H., Kim, T. K., Wei, J., Shin, J., Hui, F. K., Sair, H. I., Hager, G. D., and Fritz, J. (2019a). Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning. *Pediatric Radiology*, 49(8):1066–1070.

Yi, P. H., Lin, A., Wei, J., Yu, A. C., Sair, H. I., Hui, F. K., Hager, G. D., and Harvey, S. C. (2019b). Deep-learning-based semantic labeling for 2d mammography and comparison of complexity for machine learning tasks. *Journal of Digital Imaging*, 32(4):565–570.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks?

Yu, A. C., Mohajer, B., and Eng, J. (2022). External validation of deep learning algorithms for radiologic diagnosis: A systematic review. *Radiology: Artificial Intelligence*, 4(3).

Zebin, T. and Rezvy, S. (2020). COVID-19 detection and disease progression visualization: Deep learning on chest x-rays for classification and coarse localization. *Applied Intelligence*, 51(2):1010–1021.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. *Lecture Notes in Computer Science*, 8689:818–833.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021a). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Zhang, L., Dai, H., and Sang, Y. (2022a). Med-srnet: Gan-based medical image super-resolution via high-resolution representation learning. *Computational Intelligence and Neuroscience*, 2022.

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. *Computerized Medical Imaging and Graphics*, 77:101645.

Zhang, S., Song, G., Zang, Y., Jia, J., Wang, C., Li, C., Tian, J., Dong, D., and Zhang, Y. (2018). Non-invasive radiomics approach potentially predicts non-functioning pituitary adenomas subtypes before surgery. *European radiology*, 28(9):3692–3701.

Zhang, X., Yang, Y., Li, T., Zhang, Y., Wang, H., and Fujita, H. (2021b). CMC: A consensus multi-view clustering model for predicting alzheimer's disease progression. *Computer Methods and Programs in Biomedicine*, 199:105895.

Zhang, Z., Ji, Z., Chen, Q., Yuan, S., and Fan, W. (2022b). Joint optimization of CycleGAN and CNN classifier for detection and localization of retinal pathologies on color fundus photographs. *IEEE Journal of Biomedical and Health Informatics*, 26(1):115–126.

Zheng, H., Li, J., Liu, H., Ting, G., Yin, Q., Li, R., Liu, M., Zhang, Y., Duan, S., Li, Y., et al. (2022). Mri radiomics signature of pediatric medulloblastoma improves risk stratification beyond clinical and conventional mr imaging features. *Journal of Magnetic Resonance Imaging*.

Zheng, H., Li, J., Liu, H., Wu, C., Gui, T., Liu, M., Zhang, Y., Duan, S., Li, Y., and Wang, D. (2021). Clinical-mri radiomics enables the prediction of preoperative cerebral spinal fluid dissemination in children with medulloblastoma. *World journal of surgical oncology*, 19(1):1–10.

Zhou, L., Peng, H., Ji, Q., Li, B., Pan, L., Chen, F., Jiao, Z., Wang, Y., Huang, M., Liu, G., et al. (2021). Radiomic signatures based on multiparametric mr images for predicting ki-67 index expression in medulloblastoma. *Annals of Translational Medicine*, 9(22).

Zhu, X.-L., Shen, H.-B., Sun, H., Duan, L.-X., and Xu, Y.-Y. (2022). Improving segmentation and classification of renal tumors in small sample 3d CT images using transfer learning with convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 17(7):1303–1311.

Zong, W., Lee, J. K., Liu, C., Carver, E. N., Feldman, A. M., Janic, B., Elshaikh, M. A., Pantelic, M. V., Hearshen, D., Chetty, I. J., Movsas, B., and Wen, N. (2020). A deep dive into understanding tumor foci classification using multiparametric MRI based on convolutional neural network. *Medical Physics*, 47(9):4077–4086.